Large-scale data analysis to identify novel disease phenotypes and genes

Eevi Kaasinen, M.Sc.

Department of Medical Genetics
Haartman Institute
&
Genome-Scale Biology, Research Programs Unit
Faculty of Medicine
University of Helsinki
Finland

Helsinki Graduate Program in Biotechnology and Molecular Biology (GPBM)
/Integrative Life Science (ILS) Doctoral Program

Academic dissertation

To be publicly discussed with the permission of the Faculty of Medicine of the University of Helsinki, in Haartman Institute, Small Lecture Hall, Haartmaninkatu 3, Helsinki, on the 26th of September 2014, at 12 noon.

Helsinki 2014

Supervised by Academy Professor Lauri A. Aaltonen, M.D., Ph.D.
Department of Medical Genetics, Haartman Institute
Genome-Scale Biology, Research Programs Unit
University of Helsinki
Helsinki, Finland

Esa Pitkänen, Ph.D.
Department of Medical Genetics, Haartman Institute
Genome-Scale Biology, Research Programs Unit
University of Helsinki
Helsinki, Finland

Reviewed by Docent Marjo Kestilä, Ph.D.
Department of Chronic Disease Prevention
National Institute for Health and Welfare
Helsinki, Finland

Professor Matti Nykter, D.Sc.
Institute of Biomedical Technology
University of Tampere
Tampere, Finland

Official Opponent Docent Janna Saarela, M.D., Ph.D.
Institute for Molecular Medicine Finland
University of Helsinki
Helsinki, Finland

ISBN 978-951-51-0138-9 (paperback)
ISBN 978-951-51-0139-6 (PDF)
http://ethesis.helsinki.fi/
Unigrafia Oy
2014

Table of contents

List of original publications 5
Author's contributions 5
Abbreviations 6
Abstract 7
1 Introduction 9
2 Review of the literature 10
2.1 Human genome 10
2.1.1 Human reference genome 10
2.1.2 DNA sequence variation in human populations 10
2.2 Genetics of human disease 12
2.2.1 Disease-causing genetic changes 12
2.2.2 Genetic epidemiology 15
2.2.2.1 Isolated populations 16
2.2.3 Phenotypes relevant in this thesis 17
2.2.3.1 Heterotaxy syndrome and isomerism (I) 17
2.2.3.2 Intellectual disability (II) 18
2.2.3.3 Uterine leiomyomas (III) 19
2.2.3.4 Kaposi sarcoma (IV) 20
2.3 Genome-wide methods for studying genetic diseases 20
2.3.1 DNA microarrays 20
2.3.2 Genetic linkage analysis 21
2.3.3 Next-generation sequencing technologies 22
2.3.4 Next-generation sequencing data analysis 23
2.3.4.1 Read alignment 23
2.3.4.2 Variant calling 24
3 Aims of the study 26
4 Materials and methods 27
4.1 Study materials 27
4.1.1 Isomerism family and samples (I) 27
4.1.2 Intellectual disability family and samples (II) 27
4.1.3 Leiomyoma samples (III) 28
4.1.4 Patient data in the Finnish Cancer Registry (IV) 28

4.2 Array-based methods 28
4.2.1 SNP array data analysis and genetic mapping (I, II) 28
4.2.2 Gene expression analysis (II, III) 29
4.3 Fragment analysis 30
4.4 Sequencing methods 30
4.4.1 Whole-genome sequencing data analysis (II, III) 30
4.4.1.1 Variant calling 32
4.4.1.2 Data filtering and annotation 32
4.4.1.3 Detection of interconnected complex chromosomal rearrangements 33
4.4.1.4 Assessment of clonally related leiomyomas 34
4.4.2 PCR and Sanger sequencing (I, II, III) 34
4.5 Registry-based data analysis 34
4.5.1 Systematic clustering of patients (IV) 34
4.5.2 Estimating familiality with cluster score (IV) 35
4.6 Ethical issues 35
5 Results 36
5.1 Identification of *GDF1* mutations in right atrial isomerism (I) 36
5.2 Genetic mapping of severe intellectual disability syndrome (II) 38
5.3 Molecular genetic characteristics of uterine leiomyomas (III) 40
5.3.1 Landscape of somatic alterations and complex chromosomal rearrangements 40
5.3.2 Clonal origin of multiple tumors 42
5.4 Familial aggregation of tumor types in Finland (IV) 44
6 Discussion 46
6.1 The role of *GDF1* in isomerism and heart defects (I) 47
6.2 Candidate genes of novel severe intellectual disability syndrome (II) 49
6.3 Genetic changes in development of uterine leiomyomas (III) 50
6.4 Identification of tumor susceptibility phenotypes using registry-based data (IV) 53
7. Conclusions and future prospects 55
8. Acknowledgements 57
9. References 60

List of original publications

I **Kaasinen E**, Aittomäki K, Eronen M, Vahteristo P, Karhu A, Mecklin JP, Kajantie E, Aaltonen LA & Lehtonen R. Recessively inherited right atrial isomerism caused by mutations in *Growth/Differentiation Factor 1* (*GDF1*). *Human Molecular Genetics* 2010, 19: 2747-2753.

II **Kaasinen E***, Rahikkala E*, Koivunen P, Miettinen S, Wamelink MMC, Aavikko M, Palin K, Myllyharju J, Moilanen JS, Pajunen L, Karhu A & Aaltonen LA. Clinical characterization, genetic mapping and whole-genome sequence analysis of a novel autosomal recessive intellectual disability syndrome. *European Journal of Medical Genetics* 2014, in press, DOI: 10.1016/j.ejmg.2014.07.002.

III Mehine M*, **Kaasinen E***, Mäkinen N, Katainen R, Kämpjärvi K, Pitkänen E, Heinonen HR, Bützow R, Kilpivaara O, Kuosmanen A, Ristolainen H, Gentile M, Sjöberg J, Vahteristo P & Aaltonen LA. Characterization of uterine leiomyomas by whole-genome sequencing. *The New England Journal of Medicine* 2013, 369:43-53.

IV **Kaasinen E***, Aavikko M*, Vahteristo P, Patama T, Li Y, Saarinen S, Kilpivaara O, Pitkänen E, Knekt P, Laaksonen M, Lehtonen R, Artama M, Aaltonen LA & Pukkala E. Nationwide registry-based analysis of cancer clustering detects strong familial occurrence of Kaposi sarcoma. *PLoS ONE* 2013, 8, 1, e55209.

*Equal contribution

Author's contributions

I Performed the linkage analysis, fragment analysis in additional paraffin embedded tissue samples, literature search for candidate genes, and mutation screening. Wrote the manuscript together with other authors.

II Participated in designing the study. Performed the linkage analysis, homozygosity mapping, and whole-genome sequencing and expression data analysis. Performed and coordinated the mutation screening and functional analyses. Wrote the manuscript together with other authors.

III Participated in designing the study. Performed the whole-genome sequencing data analyses, coordinated the validation of structural variations and developed the computational method to identify interconnected complex rearrangements. Wrote the manuscript together with other authors.

IV Participated in designing the study. Calculated the familiality measures for tumor types by developing the cluster score method, coordinated the study and analyzed the clustering data. Wrote the manuscript together with other authors.

Abbreviations

bp	base pair
BWA	Burrows-Wheeler Alignment tool
cAMP	cyclic adenosine monophosphate
CCND1	cyclin D1
CCR	complex chromosomal rearrangement
cDNA	complementary DNA
CG	Complete Genomics
CHD	congenital heart defects
cM	centimorgan
CNA	copy number alteration
CNV	copy number variation
CRC	colorectal cancer
CREB	cAMP response element-binding protein
CUX1	cut-like homeobox 1
DGV	Database of Genomic Variants
DNA	deoxyribonucleic acid
DZ	dizygotic
FCR	the Finnish Cancer Registry
FDR	false discovery rate
FH	fumarate hydratase
FIMM	the Institute of Molecular Medicine Finland
GATK	the Genome Analysis Toolkit
GDF1	growth/differentiation factor 1
GRC	the Genome Reference Consortium
GWAS	genome-wide association study
HGMD	the Human Gene Mutation Database
HHV8	human herpesvirus 8
HIF-1α	hypoxia-inducible factor-1 alpha
HIV	human immunodeficiency virus
HLRCC	hereditary leiomyomatosis and renal cell cancer
HMGA1/2	high mobility group AT-hook 1/2
ID	intellectual disability
indels	insertions and deletions
IPA	Ingenuity Pathway Analysis
IQ	intelligence quotient
IRS4	insulin receptor substrate 4
KS	Kaposi sarcoma
LAI	left atrial isomerism
LOD	logarithm of odds
LOH	loss-of-heterozygosity
LPM	lateral plate mesoderm
MAF	minor allele frequency
MED12	mediator complex subunit 12
MZ	monozygotic
NCBI	National Center for Biotechnology Information
NGS	next-generation sequencing
NPR	the National Population Registry
OMIM	Online Mendelian Inheritance in Man
P4HTM	prolyl 4-hydroxylase transmembrane
PIC	personal identity code
PCD	primary ciliary dyskinesia
PCR	polymerase chain reaction
PPP	pentose phosphate pathway
RAD51B	RAD51 paralog B
RAI	right atrial isomerism
RMA	robust multi-array average
RNA	ribonucleic acid
RPI	ribose-5-phosphate isomerase
SNP	single nucleotide polymorphism
SNV	single nucleotide variation
SV	structural variation
TGFβ	transforming growth factor beta
TKT	transketolase
TSS	transcription start site
UCSC	University of California, Santa Cruz
USP4	ubiquitin specific peptidase 4
WGS	whole-genome sequencing

Abstract

Diseases can occur due to genetic changes that alter the normal function of genes. These alterations may be either inherited, thus present in every cell of an individual at birth, or acquired somatically during lifetime. In this thesis, a combination of genome-wide measurement technologies, a unique national registry of all cancer cases, and sophisticated data analysis methods were utilized to study the genetic background of human diseases. Aims of this thesis work were to efficiently analyze large quantities of epidemiological and molecular data, and to characterize new susceptibility conditions and genetic causes of human diseases.

First, unknown genetic basis of right atrial isomerism (RAI) was studied in a previously reported Finnish family with five affected siblings and healthy parents. RAI is a heterotaxy syndrome with disturbances in the left-right axis development resulting in complex heart malformations and abnormal lateralization of other thoracic and abdominal organs. Heterotaxy syndromes are associated with a few known allelic variants in humans, although studies with model organisms have identified several genes involved in the early regulation of laterality. Linkage analysis and candidate-gene approach followed by sequencing revealed two truncating mutations in *GDF1* segregating with the RAI phenotype in an autosomal recessive manner. This finding, supported by the similar phenotype of laterality defects in *Gdf1* knockout mice, provides evidence that RAI can be recessively inherited with *GDF1* as the causative gene.

Second, six clinically well-characterized patients with severe intellectual disability (ID) of unknown etiology were studied by genetic mapping and whole-genome sequencing (WGS) analysis. ID is a genetically extremely heterogeneous condition where many autosomal recessive genes are yet to be identified. In this study, autosomal recessive inheritance of severe ID was confirmed by extensive genealogy, and by linkage analysis showing the logarithm of odds score of 11 for a homozygous region at 3p22.1-3p21.1. The WGS data revealed three candidate genes, *TKT*, *P4HTM* and *USP4*, with potentially protein damaging sequence changes within the locus. The variants were present in heterozygous form with 0.3-0.7% allele frequencies in population-matched controls from Northern Finland. This study facilitates clinical and molecular diagnosis of similar patients and further research on the role of the genes in the development of severe ID.

Third, the molecular genetic landscape of uterine leioymomas was studied utilizing the most recent genome-wide technologies. Uterine leiomyomas are benign tumors that affect approximately three-quarters of all women and may cause severe symptoms including abdominal pain and excessive uterine bleeding. We sequenced the genomes of 38 leiomyomas and corresponding myometrium tissues from 30 women, and performed whole-transcriptome profiling of the same tissue specimes. Abundant complex chromosomal rearrangement events resembling the recently described chromothripsis phenomenon were detected in leiomyomas. The events had created leiomyoma-specific driver changes, and occurred sequentially in some tumors. Four mutually exclusive molecular pathways driven by alterations of *MED12*, *FH*, *HMGA2*/*HMGA1* or *COL4A5*/*COL4A*6 were identified. The clonal origin of multiple separate tumors was proven by sequence analysis. The molecular genetic characterization of uterine leiomyomas will hopefully lead to better understanding of tumor growth and personalized treatment of patients.

Fourth, a systematic search for familial aggregation of all types of cancer was performed to identify new tumor susceptibility phenotypes and families. Traditionally, information on

family relations is a prerequisite for familiality studies. We employed the entire population based data in the Finnish Cancer Registry and clustered 878,593 patients according to family name at birth, municipality of birth and tumor type. To estimate the rate of familial occurrence, a cluster score was calculated for all tumor types producing significant clusters. Known cancer predisposition syndromes displayed the highest cluster scores, and some phenotypes with largely unknown genetic background, such as Kaposi sarcoma (KS), were also highlighted. Population records verified majority of the clustered KS patients as true relatives, providing further evidence that the clustering works well in estimating familiality. The effort described in this study enabled identification of families suitable for a succeeding research on genetic basis of novel tumor predisposition phenotypes.

1 Introduction

Many diseases have a genetic cause, which can be either inherited or acquired over lifetime. Vast majority of diseases are complex, arising as a combination of genetic, environmental and life-style factors. When the inheritance of the disease is clear and caused by a single gene, such as in many congenital diseases, the pattern of inheritance can be deduced from multigenerational families. Mendelian patterns of inheritance include dominant and recessive inheritance. Complex diseases do not follow simple Mendelian patterns of inheritance, although a genetic susceptibility to develop the disease may be inherited. Cancer is an example of a complex disease that arises from mutations that accumulate somatically in the descendants of a cell over time, causing tumor growth in a specific organ of the body.

Identification of novel genetic causes of diseases is empowered by the completion of the human genome sequencing project and the advent of next-generation sequencing technologies. If a disease seems to run in families and sample materials are available from multiple affected individuals, genetic mapping followed by sequencing of candidate regions can be used to determine the disease-causing genetic changes. Genetic mapping examines co-segregation of genetic markers and disease in a family pedigree. Within the last couple of years, whole-genome sequencing has become more affordable allowing genome-wide comparison of genetic changes in multiple samples, and discovery of structural variants and complex chromosomal changes, which could not have been detected with traditional molecular approaches. The new sequencing technologies have increased our knowledge of the amount and type of variation in human populations and genetic diseases.

In this thesis work, microarrays were used to measure genome-wide genotypes of approximately 300,000 to 2 million single nucleotide polymorphisms or transcript abundance of over 35,000 genes. Next-generation sequencing technologies were used to produce millions of short sequencing reads from individual genomes which were analyzed for alterations by comparing them to the 3 billion base pairs in the human reference genome. Furthermore, over one million patient records in the Finnish Cancer Registry were analyzed for familial aggregation to estimate heritability of various tumor types. Without computational data analysis and integration methods, biological interpretation of these large-scale data would be unfeasible.

2 Review of the literature

2.1 Human genome

Hereditary information is encoded in sequences of deoxyribonucleic acid (DNA) molecules organized in 46 chromosomes in the cell nucleus, and in mitochondrial DNA. DNA molecule units, nucleotides, are composed of sugar and phosphate backbones, and of four different nucleobases, one in each nucleotide. The bases are grouped into purines [adenine (A) and guanine (G)] and pyrimidines [cytosine (C) and thymine (T)]. Purines and pyrimidines form pairs in the double-strand helix structure of DNA resolved by Watson and Crick in 1953 (Watson and Crick, 1953).

2.1.1 Human reference genome

The human reference genome is used to describe the consensus of the base pair-level composition of the haploid genome of the 22 autosomes, the X and Y chromosomes, and the mitochondrial DNA. The total length of the human reference genome is approximately 3 billion base pairs (bp) and more than 99% of this is shared between individuals (The Human Genome Project, 2014), although more and more variation has been detected in individual genomes as compared to the reference genome by utilizing new genome-wide technologies (Pang *et al.*, 2010).

Completion of the sequencing of the human genome was celebrated in April 2003 when the Build 34 version of the human reference genome was published by the International Human Genome Sequencing Consortium (The Human Genome Project, 2014). The aim of the project was to identify the nucleotide composition of the euchromatic genome with 99.99% accuracy and to make this information publicly available. The euchromatic portion of the human genome was estimated to consist of 20,000–25,000 protein coding genes (International Human Genome Sequencing Consortium, 2004). Since 2003, much work has been done to fill in gaps and refine complex sequences to produce a better consensus representation of the human genome.

The human reference genome version used in most parts of this thesis was produced by the Genome Reference Consortium (GRC) in 2009 (Build 37). The GRC is a collaborative effort of many research institutes, and it aims to produce a high quality reference assembly where any sequences longer than 500 bp are positioned into a chromosome context. The reference assembly produced by the GRC represents chromosomes as well as unlocalized, unplaced and alternate loci sequences of the human genome (Church *et al.*, 2011). These improvements in the reference assembly are needed to better account for structural diversity in human populations and to allow more accurate next-generation sequencing analysis, as described in section 2.3.4.

2.1.2 DNA sequence variation in human populations

Single nucleotide variation (SNV) is the most common variation in the human genome; it occurs when a single nucleotide in the studied DNA differs from the reference genome. The National Center for Biotechnology Information (NCBI) maintains a central repository, the Database of Short Genetic Variation (dbSNP) for single base nucleotide substitutions, for short multi-base insertions and deletions (indels), and for microsatellite repeats in the human genome as compared to the reference assembly (Kitts *et al.*, 2013). Each variation type per location in the genome is identified with a "Reference SNP" (rs) identifier. The advent of next-generation sequencing (NGS) has increased the number of rs variants from 23,653,737 in dbSNP build 131 (Mar 25, 2010) to 62,676,337 in build 138 (April 25, 2013). dbSNP manages data on the sequence context of the variant, the frequency of the polymorphism in

populations and all the relevant experimental information from the submitter, while databases such as NCBI's ClinVar database, The Human Gene Mutation Database (HGMD) (Stenson *et al.*, 2013) and Online Mendelian Inheritance in Man (OMIM) (Hamosh *et al.*, 2005) store clinical significance information of variants found in patient samples.

The International HapMap Project was started in 2002 in order to identify common single nucleotide polymorphisms (SNPs) with minor allele frequency (MAF) >5% in human populations (International HapMap Consortium, 2003). More than one million common SNPs were genotyped in the initial set of 270 samples from individuals with African, Asian and European ancestry, including 60 parent-offspring trios (International HapMap Consortium, 2005). Additional SNPs and seven more admixed populations residing in the US were genotyped in the later phases of the project (International HapMap 3 Consortium *et al.*, 2010; International HapMap Consortium *et al.*, 2007). One major goal of the HapMap Project was to identify statistically related SNPs and to create all unique haplotypes across the genotyped individuals (International HapMap Consortium, 2005). Recombination rates as well as distribution information within and between populations have also been provided for the genotyped SNPs and haplotypes by the HapMap project.

Database of Genomic Variants (DGV) was established after first reports of high prevalence of copy number variations (CNVs) (about 100 kb and greater) in the genomes of healthy individuals (Macdonald *et al.*, 2014; Sebat *et al.*, 2004). The objective of the database is to provide numerous types of structural variations (SVs), including copy number gains, duplications, insertions, inversions and complex variants observed in healthy control samples. The variant data in the database comes mostly from microarrays (44%) and sequencing studies (53%), and it is enriched for deletions and copy number losses (70%) (Macdonald *et al.*, 2014). Mapping the full spectrum of variation in an individual genome (J Craig Venter's DNA) revealed that approximately 1.2% of this genome is encompassed by indels and CNVs, 0.3% by inversions and 0.1% by SNPs. In the study of the Venter genome, the reported SVs affected 4,867 genes, some of which were linked to human disease phenotypes (Pang *et al.*, 2010).

At the time NGS technologies became available in 2008, the 1000 Genomes project was launched (1000 Genomes Project Consortium *et al.*, 2010). The aim of the project was to develop methodologies to cost-effectively and accurately detect majority of single nucleotide and structural variants with frequencies of at least 1% in 2,500 individuals from the five major population groups (European, East Asian, South Asian, West African and American ancestry). In 2012, an integrated map of SNPs, indels and larger deletions was published from the genomes of 1,092 individuals by using NGS sequencing and SNP genotyping assays. These data showed that variants with the frequency of at least 10% across all individuals were present in each of the populations studied, whereas low-frequency variants (<5%), which tend to be recent, differentiated populations by geographic origin. Finnish samples (n=93) were among the most differentiated samples showing excess of low-frequency variants reflecting an increase in the population size after a recent bottleneck (1000 Genomes Project Consortium *et al.*, 2012). Most importantly, the analysis tools and data generated by the 1000 Genomes Project facilitate NGS sequencing studies of human diseases.

2.2 Genetics of human disease

2.2.1 Disease-causing genetic changes

DNA variations in the gene coding regions of the genome may alter the function of the encoded proteins and potentially lead to a disease phenotype. SNVs causing a premature stop codon (nonsense mutation) and indels causing a change in the reading frame of a gene (frameshift mutation) are considered truncating changes, which most likely yield a non-functional protein product. Furthermore, nucleotide changes in the vicinity of splice junctions are often harmful due to splicing defects. Nonsense, frameshift and splice-site mutations are enriched among disease-causing variants, although non-synonymous missense changes may also be damaging for the protein function (Cooper and Shendure, 2011).

Delineation of disease-causing variants can be achieved by approaches such as family-based linkage analysis followed by sequencing of candidate regions and genetic validation in patients with a similar phenotype (Cooper and Shendure, 2011). Often the genetic information is insufficient to conclude causality because of low number of family and patient samples available for the study. In many cases, experimental and computational approaches can be used to assess variant function. Prediction of pathogenicity of variants, especially in the case of missense changes, relies on sequence conservation in many species, on biochemical properties of amino acids and on structural information of the encoded protein. An example of this kind of prediction tool is PolyPhen-2 (Adzhubei *et al.*, 2010). Recent NGS studies have shown individual genomes to harbor on average 150-179 loss-of-function variants (nonsense, frameshift and splice-site variants) (1000 Genomes Project Consortium *et al.*, 2012; Shen *et al.*, 2013). However, the numbers of rare (<0.5%) variants were much less, such that individuals are estimated to carry up to 20 rare loss-of-function and disease associated variants (1000 Genomes Project Consortium *et al.*, 2012). Therefore variant frequencies in the population are important to take into consideration when assessing candidate pathogenic variants (MacArthur *et al.*, 2014).

Chromosomal changes which delete, duplicate or rearrange genetic material have originally been detected by cytogenetics and used for positional cloning of single-gene disorders (Tommerup, 1993; Vissers *et al.*, 2005). Many disease phenotypes caused by a deletion are due to haploinsufficiency of dosage-sensitive genes, meaning that a single copy is not enough for the gene to function properly. Deletions that span one or more exons are estimated to account for up to 15% of all mutations in monogenic diseases (Vissers *et al.*, 2005). However, often the deletions and duplications found in single patients are novel or extremely rare, and occurrence in multiple patients is required for disease gene stratification and clinical interpretation. Centralized databases such as DECIPHER (Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources) have been established to enable data sharing especially in rare disease studies (Firth *et al.*, 2009).

Cancer is thought to arise from mutations that accumulate in descendants of a cell over time. Cancer development is an evolutionary process of acquisition of somatic variation in individual cells and selection of cells with growth advantage. A combination of advantageous mutations allows a cell to proliferate autonomously and drives a clonal expansion (Stratton *et al.*, 2009). The idea known as the two-hit hypothesis was first proposed by Knudson (1971): tumors develop when inactivating mutational events occur in both copies of a gene that has a growth-suppressive function in normal cells (a tumor suppressor gene). In particular, if the first recessive mutation is inherited in germline and the second mutation is acquired somatically, cancer occurs at a younger age than if both mutations are somatic (Knudson, 1971). In hereditary cancers, the wild type allele of a tumor suppressor gene is frequently lost

in tumors through a large-scale somatic deletion or mitotic recombination, both of which can be detected as regions of loss-of-heterozygosity (LOH) (Hansen and Cavenee, 1987). Recurrent somatic amplifications, translocations or deletions distributed along particular genes across tumors might indicate that alterations at these sites constitute tumor progression. Tumor suppressors can most easily be identified at genomic sites displaying homozygous deletions in tumors, whereas amplifications likely contain oncogenes, defined as genes whose protein product is abnormally activated increasing cell survival and proliferation (Hanahan and Weinberg, 2011; Vogelstein *et al.*, 2013).

Studies have shown that somatic point mutations are the most prevalent type of changes affecting protein coding genes, although the rate of chromosomal changes is elevated in tumors (Hanahan and Weinberg, 2011; Vogelstein *et al.*, 2013). Gains and losses of copy numbers in tumor cells are induced by genomic instability during tumor progression (Hanahan and Weinberg, 2011). NGS technologies have revealed that some copy number changes seen in cancer and developmental disorders can arise from complex chromosomal rearrangements involving at least three genomic breakpoints (Zhang *et al.*, 2009). Recently, a phenomenon termed chromothripsis was described with tens to hundreds of clustered rearrangements accompanied with focal losses which had occurred simultaneously in a single event (Figure 1). The massive remodeling event was suggested to affect one or few chromosomes and at least 2-3% of all cancers (Stephens *et al.*, 2011). Since its discovery, chromothripsis has also been shown to inactivate genes that drive tumorigenesis, such as *RB1* in retinoblastoma, or to create oncogenic fusions (McEvoy *et al.*, 2014; Parker *et al.*, 2014).

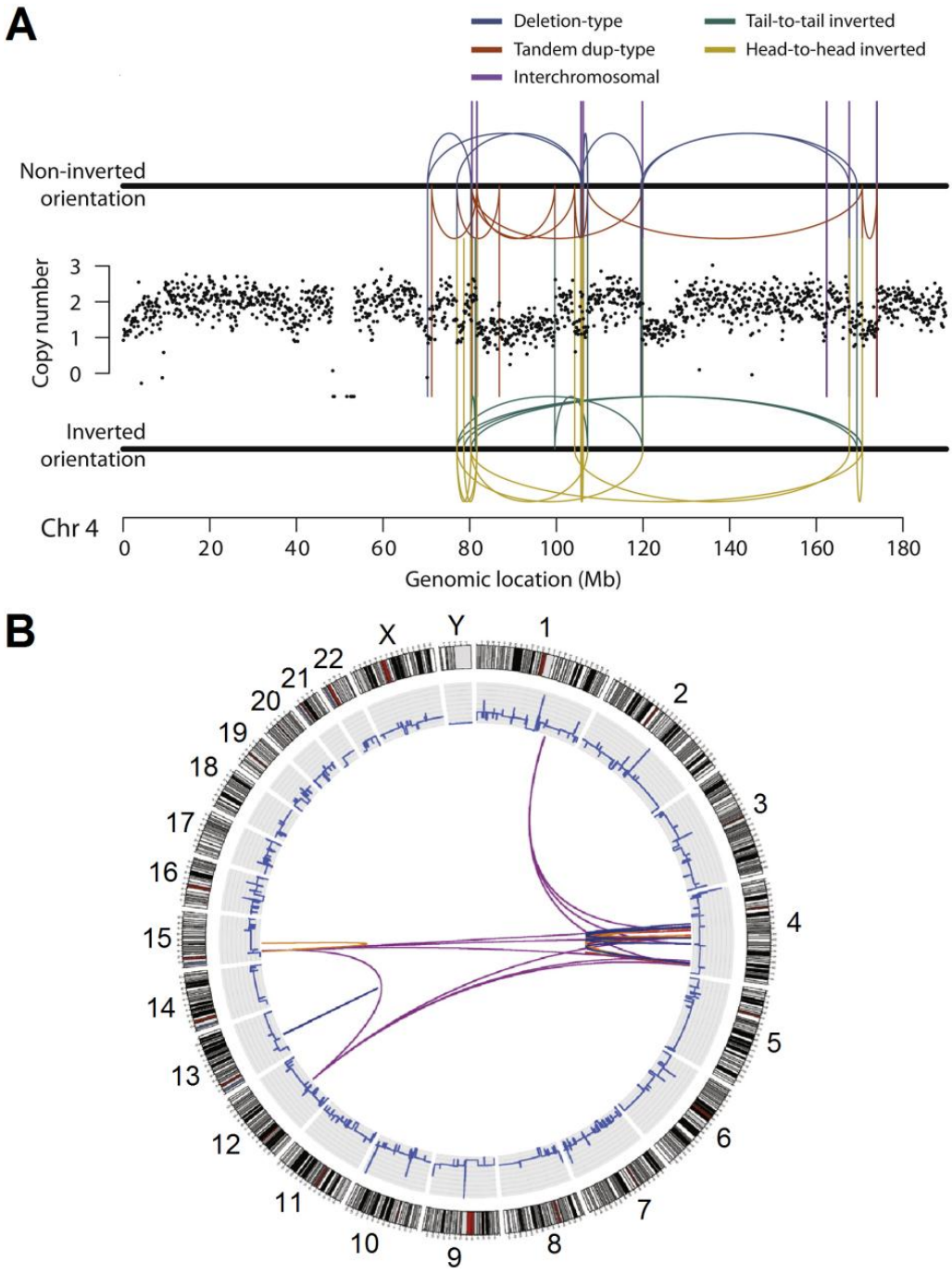

Figure 1. An example of chromothripsis in a chronic lymphatic leukemia patient sample.
(A) Clustered complex rearrangements generate copy number oscillation between 1 and 2 at genomic locations 70-170 Mb on chromosome 4. (B) Nine interchromosomal rearrangements join the chromosome 4q to chromosomes 1, 12 and 15. Chromosomes are organized circularly in the outer ring. Somatic rearrangements are shown as the colored links between the two relevant genomic locations. Modified from Stephens *et al*. (2011), copyright 2011 Elsevier Inc. Reuse permitted by Creative Commons public licence.

Many published studies have focused on disease-causing changes in the coding regions of the human genome as functional interpretation of their consequences is more evident. Genetic changes in noncoding regulatory regions of genes may disrupt transcription factor binding and increase or decrease transcriptional activity of genes. Annotations such as guanine-cytosine content, evolutionary conservation, DNase I hypersensitivity, histone modifications and distance to the nearest transcription start site (TSS) are the best indicators of functionality of noncoding regions. HGMD (the April 2012 release) contains 1,614 germline variations annotated as 'regulatory mutations' of which 75 % are located within a 2 kb distance to an annotated TSS (Ritchie *et al.*, 2014).

Genome-wide association studies (GWAS), population-based case-control studies testing for association between SNPs and a trait in hundreds or thousands of persons, have identified a large number of genetic loci for common diseases (Manolio, 2010). The associated SNP may be causative itself or genetically linked with another variation that is disease-causing. Majority of variants identified in the published GWAS studies till December 2008 were common (MAF>5%), associated with modest effect sizes (median odds ratio 1.33), and located in intronic (45%) or intergenic (43%) regions suggesting that noncoding variants have a role in the etiology of common diseases (Hindorff *et al.*, 2009).

Genetic changes which have been associated with diseases and published in the peer-reviewed literature are collected into central repositories. HGMD collects germline mutations that underlie or have an association with inherited diseases (Stenson *et al.*, 2013), whereas COSMIC gathers information on somatic mutations in cancer (Forbes *et al.*, 2011). Continual reassessment of the data is needed since NGS datasets in apparently healthy individuals, such as the 1000 Genomes project data, are bringing into question the pathogenicity of previously reported disease-causing mutations. In a recent high-throughput sequencing study of 104 unrelated individuals, 27% (122/460) of published severe recessive disease-causing mutations were found to be common polymorphisms, sequencing errors or lacking evidence for pathogenicity (Bell *et al.*, 2011). However, disease-causing variants with reduced penetrance are not uncommon, and there are many examples of modifier variants, for example, that influence penetrance of diseases (Cooper *et al.*, 2013).

2.2.2 Genetic epidemiology

Combination of two disciplines, genetics and epidemiology, has been considered as a distinct entity, genetic epidemiology since mid-1980s. Genetic epidemiology "focuses on the role of genetic factors and their interaction with environmental factors in the occurrence of disease in human populations", although different views on the scope of genetic epidemiology exist (Khoury *et al.*, 1993). Some of the definitions restrict genetic epidemiology mainly to the study of familial aggregation, whereas other definitions emphasize the joint analysis of genetic and environmental factors in disease etiology. The broad goal of genetic epidemiology is to understand genetic background of diseases at population level, and to work towards disease control and prevention (Khoury *et al.*, 1993).

Research strategies of genetic epidemiology comprise of population and familial aggregations studies. Population studies try to determine the distribution of diseases and genetic traits in a population and the role of genetic factors in disease etiology (Khoury *et al.*, 1993). According to King *et al*. (1984), studies of familial aggregation address three main questions: does a disease cluster in families; is familial clustering caused by common environmental exposure, inherited susceptibility or culturally transmitted risk factors; and, finally, what is the model of inheritance. Genealogical data in the Utah Population Database, Swedish Family-Cancer Database and Icelandic Cancer Registry have been employed in the largest familial

aggregation studies of cancer (Albright F Ph *et al.*, 2012; Amundadottir *et al.*, 2004; Czene *et al.*, 2002; Goldgar *et al.*, 1994).

Epidemiological measures of relative risk are used to examine association between exposure and disease occurrence. For example, relative risk for familial aggregation can be calculated by comparing disease frequency in relatives of affected individuals with disease frequency in relatives of unaffected individuals or with general population (Khoury *et al.*, 1993). Three types of measures of relative risk, namely risk ratio, rate ratio and odds ratio, can be calculated. In particular, a relative risk value greater than 1.0 indicates an increased risk for the disease among individuals in the exposed group. If no direct data is available for the comparison group, proportional incidence measures can be calculated by dividing observed number of cases (O) by the expected number (E). Population sub-groups (strata) may have marked differences in disease occurrence, which is adjusted by calculating stratum-specific measures (Santos, 1999).

In addition to familial aggregation studies, contribution of genetic and environmental factors in disease etiology can be estimated with twin studies (King *et al.*, 1984). As twins share many environmental exposures and cultural risk factors, monozygotic (MZ) and dizygotic (DZ) twins should be 100% and 50% concordant for the disease, respectively, if the disease was completely genetically determined. A twin study based on data from Swedish, Danish and Finnish twin registries found concordances less than 10% for many of the cancer sites examined (Lichtenstein *et al.*, 2000). Most of the concordances were greater in MZ than in DZ twins, supporting the existence of a genetic factor. Heritability was estimated using a statistical model in which phenotypic variance in twins was divided into hereditary, shared environment and non-shared environment components. The highest hereditary effects (26%-42%) were observed for stomach, colon/rectum, breast, prostate and lung cancers. The risk of getting stomach cancer, for example, was estimated to be accounted for by 28% of hereditary, 10% of shared environment and 62% of non-shared environment effects (Lichtenstein *et al.*, 2000).

The difficulty in multifactorial models of inheritance is to specify the effects of various non-genetic risk factors, which are selected based on tractability rather than understanding of underlying biological mechanism in disease etiology. Unfortunately, increase in genomic information has not yet been followed by development of methodologies to study joint effects of genes and environment (Khoury *et al.*, 2011).

2.2.2.1 Isolated populations

Families with multiple affected individuals are important in finding evidence for genetic factors in diseases. Affected individuals derived from heterogeneous populations may provide limited power for association and linkage studies because the phenotype of individuals with genetic predisposition can vary under different environmental conditions, and there can be different genetic causes. Therefore homogeneous populations, especially isolated populations are often utilized in the mapping of disease genes (Heutink and Oostra, 2002).

The relatively homogenous Finnish population, encompassing around 5.4 million people, has been successfully utilized in the gene mapping of Mendelian disorders. The Finnish population has traditionally been divided into early-settlement and late-settlement regions. The southwest and southeast coastal regions, recognized as the early settlement, were inhabited first in the history of the Finnish population. In the sixteenth century, the population in the small southeastern area started to expand to the inland areas of Finland resulting in the late settlement (Peltonen *et al.*, 1999b). Ten distinct subpopulations have been identified

among early- and late-settlement regions with high-density SNP genotyping, suggesting multiple bottlenecks and population growth by expansion in these wide inland areas of Finland. The youngest subisolates in the north-east of Finland were shown to display higher homozygosity across the genome as compared to the early-settlement population, and high linkage disequilibrium as compared to other isolates worldwide (Jakkula *et al.*, 2008). The small number of founders in isolated populations allows enrichment of diseases caused by single-origin mutations, which can be studied in consanguineous patients. In Finland, the information on relatedness of patients can be derived from genealogical records in the parish registries and in the National Population Registry (NPR) since 1580 (Peltonen *et al.*, 1999a).

2.2.3 Phenotypes relevant in this thesis

In this thesis, both monogenic and complex diseases were studied by means of genetics and epidemiology. Monogenic diseases with full penetrance are assumed to obey the dominant or recessive Mendelian patterns of inheritance, while factors such as incomplete penetrance, age at onset and phenocopies complicate identification of disease segregation. Monogenic diseases might also arise in a patient due to a new, *de novo* mutation. Complex diseases are multifactorial diseases that arise from combination of genetic, environmental and life-style factors. Monogenic inheritance might also be oversimplified in many Mendelian traits because of combinatorial effects of multiple genetic factors on a single patient's phenotype (Badano and Katsanis, 2002). Diseases that follow simple Mendelian patterns of inheritance tend to be rare, whereas many common diseases, such as cancer, are complex.

Sporadic cancers can often have the same underlying genetic defects that have been identified in Mendelian cancer predisposition families. Cancer predisposition is a rare condition that is estimated to account for approximately 3% of cancers, although 40% of the cancer predisposition genes are found to be mutated also in sporadic tumors (Rahman, 2014). For example, high-penetrance germline mutations in genes such as *MLH1*, *MSH2*, *APC* and *MYH* have been identified in large cancer syndrome families (MIM #609310, #120435, #175100, and #608456, respectively), and the predisposed individuals account only for up to 5% of all colorectal cancer (CRC) cases (Aaltonen *et al.*, 2007). However, the same genes and signaling pathways are also central in the development of sporadic colorectal tumors.

2.2.3.1 Heterotaxy syndrome and isomerism (I)

During the development of the vertebrate embryo, there is an initial event breaking the bilateral symmetry and a consequent establishment of left-right information with side-specific gene expression. Studies on various vertebrate model organisms have revealed similarities as well as divergences in left-right axis determination pathways that seem to converge on a node, which is a transient structure formed during the cell-migrating phase of a developing embryo (gastrulation) (Raya and Belmonte, 2006). Cells in the node are monociliated and the clockwise rotation of the cilia generates a leftward flow of extracellular fluid known as nodal flow, which is suggested to initiate the left-right asymmetry (Nonaka *et al.*, 1998; Okada *et al.*, 1999).

Defects in the left-right axis specification during embryogenesis result in abnormal arrangement of asymmetrical structures in the human body. The normal arrangement is designated as *situs solitus*, while the two types of abnormal arrangement are *situs inversus*, a complete and mirror-imaged reversal of asymmetrical organs, and *situs ambiguus*, a combination of *situs solitus* and *situs inversus*. *Situs ambiguus* is also called as heterotaxy syndrome, isomerism sequence, and Ivemark, asplenia or polyasplenia syndrome. Heterotaxy syndrome is considered whenever asymmetrical organs are not in their usual or mirror-imaged arrangement (Cohen *et al.*, 2007).

Heterotaxy syndrome is subdivided into right atrial isomerism (RAI) and left atrial isomerism (LAI), whose characterization can be based on atrial appendages that have right-sided or left-sided morphology on both sides of the heart, respectively (Cohen *et al.*, 2007). In addition to complex cardiac malformations, right isomerism is often associated with asplenia, whereas left isomerism is associated with polysplenia (Ivemark, 1955). In addition, the lungs can be bilaterally trilobular or bilobular, and stomach and liver can be located in abnormal positions in the abdomen (Cohen *et al.*, 2007). Patients with RAI or LAI may have different combinations of cardiac and extra-cardiac anomalities, although cardiac anomalities mainly dictate the long-term outcome of the patients. RAI is considered to have worse prognosis as compared to LAI because of more severe cardiac defects (Lim *et al.*, 2005).

Familial heterotaxy has been reported with autosomal dominant, recessive and X-linked inheritance. Affected siblings within a heterotaxy family may have different situs variants, which illustrates the heterogeneity of laterality defects and the complexity of left-right axis development (Casey, 1998). Mutations in the *ZIC3* gene are reported to cause heterotaxy in multiple families showing X-linked inheritance (MIM #306955). In a study by Gebbia *et al*. (1997), affected males with mutations in *ZIC3* had *situs ambiguus*, whereas heterozygous females in four of the families were anatomically normal. In one family, three out of nine heterozygous females had *situs inversus*, but the rest of the heterozygous females were unaffected (Gebbia *et al.*, 1997).

Primary ciliary dyskinesia (PCD) is a laterality defect that arises as a result of structurally and functionally defective cilia. PCD patients have often respiratory and upper airway symptoms, male infertility, and *situs inversus (*Noone *et al.*, 2004*)*. A small portion of PCD patients have heterotaxy, and cardiac and/or vascular abnormalities which are associated with mutations in the genes that code for outer dynein arm components of cilia, such as *DNAI1* and *DNAH5* (Kennedy *et al.*, 2007). Variations in the *CFC1* (MIM #605376), *ACVR2B* (MIM #613751), *NODAL* (MIM #270100), *CCDC11* (MIM #614779) and *LEFTY2* genes, as well as a translocation breakpoint at 6q21 have been reported in patients with autosomal heterotaxy or left-right axis malformations (Bamford *et al.*, 2000; Kato *et al.*, 1996; Kosaki *et al.*, 1999a; Kosaki *et al.*, 1999b; Mohapatra *et al.*, 2009; Peeters *et al.*, 2001; Perles *et al.*, 2012). Of these, only *CCDC11* has been associated with autosomal recessive inheritance of heterotaxy.

2.2.3.2 Intellectual disability (II)

Intellectual disability (ID), also known as mental retardation or early-onset cognitive impairment, refers to a condition with delayed development and reduced ability to cope independently. Severity of retardation can be assessed in early childhood with tests measuring verbal and motor performance. Diagnosis of ID is based on an intelligence quotient (IQ), and individuals with IQ<50 are considered to have a severe form of ID (Ropers, 2010). Conventionally, the disease is further categorized as syndromic ID, if other abnormalities exist beside cognitive impairment.

ID has several environmental risk factors, including malnutrition, maternal transmission of infectious diseases and fetal alcohol exposure. In developed countries, severe form of ID is mostly genetically determined. Chromosomal changes are estimated to account for ~25% of all patients with ID, and X-linked gene defects for ~10% of males with ID. Down syndrome (#190685) caused by trisomy of chromosome 21 is the most frequent genetic form of ID. Cytogenetically visible deletions have been identified in a number of ID syndromes with recognizable clinical features, such as Prader-Willi syndrome (MIM #176270) and Angelman syndrome (MIM #105830) (Ropers, 2010). Both of these syndromes have multiple genetic and epigenetic etiologies but the majority of the cases (~70%) have *de novo* interstitial

deletion of 15q11-q13 (Horsthemke and Wagstaff, 2008). The *FMR1* gene defect on X chromosome, which underlies the fragile X syndrome, is the second most frequent genetic cause of ID (MIM #300624) (Rousseau *et al.*, 1995). Many inborn errors of metabolism, such as phenylketonuria (MIM #261600), are recessively inherited ID disorders with mutations in individual genes that code for enzymes (Ropers, 2010). Genetic causes of X-linked ID are known better than those of autosomal recessive ID, although autosomal recessive ID is considered to be the most common form of ID in populations with high rate of parental consanguinity (Musante and Ropers, 2014). Recently, NGS technologies have revealed novel genes for autosomal dominant *de novo* ID and autosomal recessive ID (de Ligt *et al.*, 2012; Najmabadi *et al.*, 2011). The large number of known ID genes demonstrates that the etiology of ID is genetically heterogeneous.

2.2.3.3 Uterine leiomyomas (III)

Uterine leiomyomas (also known as fibroids) are benign smooth muscle tumors with an estimated prevalence of 70-80% among women of reproductive age (Catherino *et al.*, 2011; Cramer and Patel, 1990). Leiomyomas can cause a variety of symptoms, including abdominal pain and excessive uterine bleeding, determined by the size and location of the tumor. Severe symptoms develop in 15-30% of women (Catherino *et al.*, 2011). One large predominant tumor or many tumors of varying size can grow in a single uterus. The ovarian hormones estrogen and progesterone are essential for leiomyoma growth (Bulun, 2013). Uterine leiomyomas rarely develop into malignant cancer, and they are the most common medical reason for hysterectomy (Cramer and Patel, 1990; Leibsohn *et al.*, 1990).

Epidemiological studies have suggested African-American ethnicity, obesity, age and nulliparity to increase the individual risk for leiomyomas (Flake *et al.*, 2003). Many of the risk factors have been proposed to have an effect on estrogen and progesterone levels, which in turn increase the likelihood of somatic mutations and tumor formation (Rein, 2000). Existence of inherited genetic predisposition has been suggested based on higher incidence of leiomyomas among African-American women (Marshall *et al.*, 1997) and higher concordance for leiomyomas in MZ than in DZ twin pairs (Luoto *et al.*, 2000).

Hereditary leiomyomatosis and renal cell cancer (HLRCC) syndrome (MIM #150800) was identified in several families with multiple cutaneous and uterine leiomyomas, and the predisposition locus was localized to chromosome 1q42.3-q43 (Alam *et al.*, 2001; Launonen *et al.*, 2001). Subsequently, heterozygous germline mutations were identified in the *fumarate hydratase* (*FH*) gene (Tomlinson *et al.*, 2002). *FH* has been shown to be a classical tumor suppressor gene inactivated by loss of the wild type allele in patients' tumors (Alam *et al.*, 2001; Kiuru *et al.*, 2001; Launonen *et al.*, 2001). Somatic bi-allelic inactivation of *FH* has also been reported in a small subset (1.3%) of nonsyndromic leiomyomas (Lehtonen *et al.*, 2004).

The most common cytogenetic changes detected in leiomyomas involve translocations between chromosomes 12 and 14 [t(12;14) (q14-q15;q23-q24)], deletions on chromosome 7 [del(7)(q22q32)] and abnormalities at 6p21 (Flake *et al.*, 2003). These rearrangements target the *high mobility group AT-hook* genes, *HMGA2* at 12q14-15 and *HMGA1* at 6p21 (Ligon and Morton, 2000), and *RAD51 paralog B* (*RAD51B*) at 14q24 (Ingraham *et al.*, 1999). These nonrandom chromosomal changes are detected in approximately 40-50% of leiomyomas (Flake *et al.*, 2003). However, the most frequent somatic mutations are observed in the *mediator complex subunit 12* (*MED12*) gene affected mainly by point mutations and, to a lesser extent, with indels and splice site defects. *MED12* mutations are found in 70% of leiomyomas, regardless of ethnic background of patients (Mäkinen *et al.*, 2011a; Mäkinen *et*

al., 2011b). The mutations in *MED12* are specific to exon 2 and have not been detected to co-occur with *HMGA2* alterations (Mäkinen *et al.*, 2011b; Markowski *et al.*, 2012).

2.2.3.4 Kaposi sarcoma (IV)

As first described by Moritz Kaposi (1872), Kaposi sarcoma (KS) appears as brownish red to bluish red skin lesions, although tumors can also develop in other organs, such as oral cavity, lymph nodes and gastrointestinal tract. At the start of the AIDS epidemic, KS was found to be enriched among homosexual men with human immunodeficiency virus (HIV) infection (Safai *et al.*, 1985), suggesting an infectious etiology. Human herpesvirus 8 (HHV8), also known as Kaposi sarcoma herpesvirus, was identified as the cause for the disease in 1994 (Chang *et al.*, 1994). HHV8 is primarily transmitted through saliva, and it can establish life-long latency in blood cells of a human host after initial infection. Lytic reactivation causes release and dissemination of progeny viruses that can subsequently infect dermal cells and cause disease manifestation in the host. Reduced immunity is essential for HHV8 reactivation (IARC Working Group on the Evaluation of Carcinogenic Risks to Humans, 2012).

KS is one of the most important viral-induced cancers, although its incidence varies strongly in different populations. KS has the highest incidence in sub-Saharan Africa, mostly due to the spread of HIV and high seroprevalence rate (>40%) of HHV8 (Mesri *et al.*, 2010). In addition to HIV infected individuals, immunosuppressed patients after organ transplantation are at higher risk for KS (Mbulaiteye and Engels, 2006). HHV8 infection alone is not sufficient for KS development, and genetic predisposition to KS has been reported in few isolated childhood cases with recessive loss-of-function mutations in *IFNGR1*, *WAS*, *STIM1* and *TNFRSF4* (Byun *et al.*, 2010; Byun *et al.*, 2013; Camcioglu *et al.*, 2004; Picard *et al.*, 2006).

2.3 Genome-wide methods for studying genetic diseases

2.3.1 DNA microarrays

Microarray based whole-genome SNP genotyping is commonly used in genetic mapping of diseases. Illumina's (Illumina Inc., San Diego, CA, USA) genotyping process with Infinium assay consists of four steps: (i) whole-genome amplification, (ii) hybridization to an oligonucleotide probe array, (iii) array-based primer extension SNP scoring, and (iv) signal amplification/staining (Gunderson *et al.*, 2005). Amplified genomic DNA is hybridized to an array that contains synthetic oligonucleotide probes of 75 bases, of which 50 bases are for target capturing and 25 bases for decoding. Illumina's BeadChip arrays are manufactured using oligonucleotide probes immobilized on microscopic beads, which are self-assembled into wells on arrays. After random assembly of beads into wells, decoding is performed to identify the location of each bead type (Steemers and Gunderson, 2005). In a single-base extension assay, one bead type is designed to capture each SNP locus. The probe sequences are designed to capture genomic DNA so that the 3' terminal base of a probe sequence is the base before a SNP locus, and the primer extension with either of the two hapten-labelled dideoxynucleotides (one for C and G, and one for A and T) corresponds to the complementary base on the genomic sequence. Probes with hapten-labelled nucleotides at the 3' end are labelled with either Cy3 or Cy5 fluorescent dye, and the signal is amplified with immunohistochemistry-based methods for the image scanning of the array (Steemers *et al.*, 2006). Infinium assay generates two intensity values (X, Y) for each SNP; one for each fluorescent dye corresponding to the two alleles (A, B) of the SNP. Illumina's BeadStudio, and more recently GenomeStudio, analysis software (Illumina Inc.) read and normalize the intensity data. The normalization algorithm adjusts the dye-dependent background and scales the intensity values to ~1 on a sub-bead pool level that is a set of beads manufactured together

(Peiffer *et al.*, 2006). This normalization process is needed to generate accurate genotyping calls for downstream analyses.

Gene expression profiling using microarrays allows quantitative analysis of tens of thousands of ribonucleic acid (RNA) molecules simultaneously (Lockhart *et al.*, 1996). For example, the Affymetrix GeneChip Human Exon Array (Affymetrix, Santa Clara, CA, USA) contains oligonucleotide probes which are complementary for exonic sequencies over 25 bp in length for a variety of annotated human, mouse or rat complementary DNAs (cDNAs). With this probe selection, transcript diversity, such as novel splice-variants, can be detected. Labeled RNA samples are hybridized on arrays that contain synthesized probes in known locations, and signal intensities of the hybridization reactions are measured (Lockhart *et al.*, 1996). Typically each molecule of interest is represented by a probeset that contains 11-20 probes. Technical noise and probe-specific affinities affect signal intensities, which need to be normalized across arrays before comparison of expression levels. Robust multi-array average (RMA) method (i) corrects for background signal, (ii) uses quantile normalization, and (iii) performs the median polish procedure to the $\log_2$ transformed normalized probe intensities, selected to contain only perfect match probes (Irizarry *et al.*, 2003). Median polish proceduce is used to protect against outliers by smoothening probe signal intensities between arrays of the same experiment and probes of the same probeset. The method operates on matrices where rows represent different arrays of an experiment and columns represent probes of a probeset. Median values of rows and columns are repeatedly substracted from probe intensity values in a matrix until the matrix stabilizes or a limit on the number of iterations is reached (Holder *et al.*, 2001). The final values in a matrix after iterations are substracted from the original probe intensity signals. The average over probe signal intensities of a probeset is then used as a measure of the expression of a molecule of interest. As the information on the human genome and transriptome has evolved since the design of the GeneChip arrays, reassignment of probes into probesets representing known genes, transcripts and exons can be done with customized chip description files (CDF files) during signal processing (Dai *et al.*, 2005).

2.3.2 Genetic linkage analysis

Sequential alleles which are inherited together in a chromosome from a parent to an offspring are genetically linked. Alleles in the same parental haplotype are separated only through meiotic recombination, in which paired chromosomes exchange homologous DNA sequences. The frequency of these recombination events between chromosomal loci is not equivalent to their physical distance in basepairs which is why another distance metric, centimorgan (cM), is used to describe genetic linkage. One centimorgan is equal to 1% change of a recombination event between two loci in a chromosome. An early linkage map of human chromosomes was created in the 1980s using restriction fragment length polymorphisms to represent the order of loci in centimorgans (Botstein *et al.*, 1980). Nowadays, SNPs are commonly used as markers in linkage maps.

Genotypes of polymorphic markers whose map positions are known are used in family-based linkage analysis to test for co-inheritance with the disease. Depending on an assumed disease inheritance model, affected members of a family should have the same combination of alleles in both chromosomes (recessive inheritance) or in one chromosome (dominant or X-linked inheritance) within the genomic region causing the disease (trait locus).

To calculate statistics for linkage between two loci, Morton (Morton, 1955) introduced the logarithm of odds (LOD) score method. The probability of meiotic recombination between two loci is called the recombination fraction, theta (θ). The recombination fraction never

exceeds the value 0.5, which represents a situation where two loci are segregating independently, in other words, they are located on separate chromosomes or far apart from each other in the same chromosome. The LOD score method compares, if the likelihood L(θ), where θ is any recombination fraction between 0 and 0.5, for two loci is higher than the null hypothesis, L(θ=0.5) for no linkage. The LOD score is the logarithm of the likelihood ratio:

$Z(\theta) = \log_{10}[L(\theta) / L(\theta=0.5)]$.

In order to calculate the likelihood L(θ), the numbers of recombinant (r) and nonrecombinant (s) family members are determined. The LOD score equation then is

$Z(\theta) = \log_{10} [(1 - \theta)^s \theta^r / (0.5)^{r+s}]$.

The haplotypes of parental chromosomes should be known once the numbers of recombinant and nonrecombinant cases are calculated. However, there might be several equally likely phases of parental chromosomes, so the overall likelihood L(θ) becomes a sum of likelihoods. For example, if there are two equally likely parental phases, the LOD score equation is:

$Z(\theta) = \log_{10} \{[½ (1 - \theta)^{s1} \theta^{r1} + ½ (1 - \theta)^{s2} \theta^{r2}] / [½ (0.5)^{r1+s1} + ½ (0.5)^{r2+s2}]\}$

The value of LOD score is determined at a recombination fraction θ that maximizes the log-likelihood. The LOD score >3 is traditionally used as a criterion for linkage in autosomes (Morton, 1955). This represents the likelihood ratio of 1000:1, but in small pedigrees where the number of family members is limited, the threshold cannot usually be reached.

Linkage analysis for more than two loci is performed with multipoint linkage methods which have more power to detect linked chromosomal regions. First, the inheritance pattern of marker genotypes at each locus is determined for a pedigree. Second, the likelihood ratio for the inheritance pattern is calculated under the hypothesis that the locus is a trait locus versus the hypothesis that a disease is unlinked to the locus (Kruglyak *et al.*, 1996). In parametric linkage analysis, the scoring depends on the assumed penetrance values and allele frequencies for the tested locus.

Multipoint linkage analysis is a computationally intensive task, and different approaches exist. These include the Elston-Stewart algorithm (Elston and Stewart, 1971), Lander-Green algorithm (Lander and Green, 1987) and Markov-Chain Monte-Carlo method (Guo and Thompson, 1992), which serve as a basis for more sophisticated linkage analysis programs. The Lander-Green and Markov-Chain Monte-Carlo approaches are implemented in MERLIN (Abecasis *et al.*, 2002) and SimWalk2 (Sobel and Lange, 1996), respectively. These were the linkage analysis programs used in this thesis.

2.3.3 Next-generation sequencing technologies

The most widely used NGS technology, which is well suited for variant discovery in human genome resequencing studies, was developed by Illumina (Metzker, 2010). The genomic DNA is first randomly sheared into fragments of a certain size distribution, adaptors are ligated to the ends of target fragments, and polymerase chain reaction (PCR) amplification is performed. These target libraries are immobilized onto a glass slide (flowcell), where each DNA molecule is clonally amplified to form millions of spatially separate clusters that undergo the sequencing reaction. Sequencing is performed in repeated cycles of single-base extension by adding all four nucleotides labeled with different dye and using four-color imaging. Each nucleotide signal is a consensus of the identical templates in the clusters, and a measure of uncertainty (a quality value) for each base is applied by a base-calling algorithm. In the paired-end setting, opposite strands from both ends of DNA fragments are sequenced

(Bentley *et al.*, 2008). Typically, millions of DNA fragments of about 300-400 bp on average are sequenced in parallel with 100 bp read length from both ends in the whole-genome sequencing (WGS) applications of the Illumina technology.

In addition to WGS, the Illumina NGS technology can be used to study whole-transcriptome (RNA sequencing), or targeted genomic regions, such as protein coding regions (exome sequencing) or DNA binding sites of proteins (ChIP-sequencing) (Metzker, 2010).

Another NGS technology, developed by Complete Genomics (CG) (Complete Genomics Inc., Mountain View, CA, USA) uses DNA nanoballs to obtain tandem copies of fragmented genomic DNA. Each DNA nanoball contains adapters inserted in a fragmented DNA molecule which is circularized. Sequencing is performed by adding fluorescent-labeled probes that are anchored by the adapter sequences (Drmanac *et al.*, 2010). This combinatorial probe-anchor ligation technology avoids accumulation of errors in contrast to sequencing by synthesis reaction used in the Illumina technology. CG sequencing is available only as a service that includes all sequencing data processing, and customers are provided with annotated variant calls.

2.3.4 Next-generation sequencing data analysis

2.3.4.1 Read alignment

Human genome resequencing studies start with the alignment of raw sequencing reads to the known reference genome. Aligners such as ELAND and Burrows-Wheeler Alignment tool (BWA) allow the mapping of a large volume of short (~35-100 bp) sequencing reads produced with the Illumina NGS technology (Bentley *et al.*, 2008; Li and Durbin, 2009). The reference genome is scanned to find best matches for each read utilizing string matching algorithms, and allowing a specified number of mismatches and indels as compared to the reference sequence. The memory and time requirements for scanning millions of reads through the whole reference sequence are manageable using Burrows-Wheeler transform as implemented in BWA. The per-base quality values are utilized to weight the contribution of each base call to the alignment of a sequencing read. When processing paired-end data, the two reads from the opposite ends of a sequenced fragment can be aligned together to find the most likely mapping position in the reference genome (Li and Durbin, 2009). A measure of confidence, mapping quality, is reported for each read alignment by aligners, and alignments are output in the standard SAM (Sequence Alignment/Map) or BAM (Binary Alignment/Map) format (Li *et al.*, 2008; Li and Durbin, 2009).

About 50 % of the human genome comprises of repetitive DNA sequences. These regions cause most of the problems in the alignment of short sequencing reads. If a read is shorter than a repetitive DNA sequence, it can map equally well to multiple locations in the genome. Sometimes a uniquely mapping read-pair can aid the alignment to the correct position. Aligners have different strategies to handle reads that map to multiple locations but correct interpretation of variant calls within repetitive regions remains challenging without increasing the length of sequencing reads (Treangen and Salzberg, 2011).

Reads spanning indels are also difficult to align with the fast string matching algorithms. Mapping of these reads may result in misalignments where many bases of the spanning reads show mismatches, and may be falsely considered as SNVs in the variant calling. Therefore initial alignments can be refined using local realignment of the reads around predicted indel positions utilizing multiple sequence alignment methods. In a well-characterized sample of the 1000 Genomes Project, 15% of reads spanning known indel sites were initially misaligned.

Of these, local realignment was shown to eliminate 1.8 million loci with significant number of mismatches (DePristo *et al.*, 2011).

The PCR amplification step in the construction of Illumina sequencing libraries may cause some DNA fragments to be overrepresented and thus, many nonindependent sequencing reads to span the same bases in the initial alignment. Variant calling, as described next, expects that each read represent a unique fragment of DNA of the original sample. Therefore, paired-end reads that have exactly the same coordinates in the alignment are marked as duplicates, and only the read in the set with the best mapping quality is used in the variant calling (Day-Williams and Zeggini, 2011).

2.3.4.2 Variant calling

Variant calling from high-throughput resequencing data depends on the proportion of non-reference alleles in the total number of aligned reads at a particular genomic site. Thus, sequencing depth has a major influence on accuracy and sensitivity of variant detection. An average read depth of at least 30 throughout the genome is recommended for comprehensive detection of heterozygous variants in the Illumina NGS data (Bentley *et al.*, 2008). The type of study project and NGS data direct the selection of variant analysis approaches. In Mendelian disease studies, for example, higher sensitivity is required with the cost of high error rate, whereas in population data, higher specificity is favored. Here, a set of programs for detection of different types of variants, including SNVs, indels, SVs and CNAs from aligned NGS reads are presented, as they were used in the WGS data analysis of this thesis.

The current NGS technologies can produce high rates of variant artifacts due to base calling and alignment errors, which can be controlled for in variant calling by requiring non-reference bases to have high base qualities, to account high percentage of the aligned reads, and to be present in reads that are mapped on both plus and minus strands of the genome (Mokry *et al.*, 2010). A framework for NGS data processing, and for SNV and indel calling is offered by the Genome Analysis Toolkit (GATK) which is designed for large-scale sequencing data, such as for the 1000 Genomes Project data. The GATK Unified Genotyper is a sensitive variant caller using the Bayesian genotype likelihood model that determines a phred-scaled confidence score for each variant (McKenna *et al.*, 2010). A commonly used verification measure for specificity of SNV calling in germline data is the transition to transversion ratio which is around 2.1 for the whole genome (DePristo *et al.*, 2011). When analyzing somatic mutations from tumors, different variant calling thresholds are required because tumor heterogeneity and purity can affect the proportion of non-reference alleles in the aligned reads. The somatic analysis tool VarScan 2 processes data from both tumor and normal sample simultaneously, and compares SNV or indel calls pairwise at each position that meets the requirement for a user-specified minimum read depth. Significance of allele frequency difference in the tumor and normal sample is calculated using Fisher's Exact test, and statistically significant variants present in the tumor and not in the corresponding normal tissue are classified as somatic (Koboldt *et al.*, 2012).

BreakDancer is an algorithm to call SVs including deletions, insertions, inversions, and translocations from paired-end NGS reads. BreakDancer detects SVs based on clusters of discordantly aligned read-pairs which are separated with a distance differing from expected insert size distribution, or which have abnormal mapping orientations, in other words, read-pairs are not on the opposite strands of the reference genome as expected with the Illumina NGS technology (Figure 2) (Chen *et al.*, 2009). BreakDancer can detect large SVs with similar or higher sensitivity than other methods that utilize discordant read-pairs in variant calling (Chen *et al.*, 2009; Pabinger *et al.*, 2014). However, complementary methods are

needed to call copy number alterations (CNAs) from the NGS data (Mills *et al.*, 2011; Pabinger *et al.*, 2014). Read-depth analysis is commonly used in combination with read-pair analysis to detect large CNAs (≥1 kb). VarScan 2, for example, produces normalized and $\log_2$ transformed tumor versus normal depth of coverage ratios along user-specified contiguous chromosomal regions, which are further processed with circular binary segmentation to detect regions with equal copy numbers (Koboldt *et al.*, 2012; Venkatraman and Olshen, 2007).

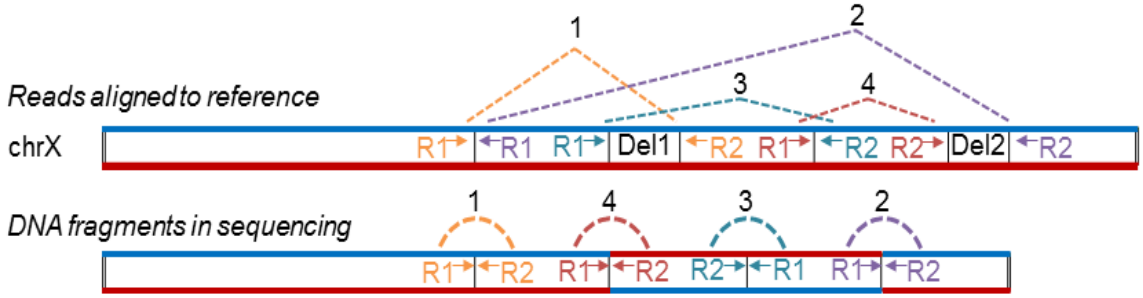

Figure 2. Schematic presentation of somatic structural variation (SV) and copy number alteration (CNA) calls detected by BreakDancer and circular binary segmentation, respectively, from the paired-end sequencing reads of a tumor in study III. Upper bar displays the SV and CNA calls on the reference chromosome X; lower bar represents the resolved structure of chromosome X of the sequenced sample. The two reads from the opposite ends of a sequenced DNA fragment, read1 (R1) and read2 (R2) are called discordant, if they are not aligned on the opposite direction (arrows), or if the distance between the reads differs from the expected insert size (distribution of distances between the 3' ends of R1 and R2 in the sequenced DNA fragments). Here, BreakDancer detected four clusters of discordant read-pairs. SVs 1 and 3 are called as deletions, because reads align to the reference genome at longer distance from each other than expected, although the fragments between the discordant read-pairs are not truly deleted. SVs 2 and 4 are called as inversions, because reads align in the same direction. CNA calls from the read depth analysis revealed two deletions (Del1 and Del2). Overall inspection of the calls reveals that the sample has interconnected complex chromosomal rearrangements on chromosome X.

The effects of alignment errors on variant calls can be reduced by multiple sequence alignment in regions of likely variation, including SNVs, indels and SVs – a procedure which is implemented in CG's variant calling procedure. SVs in the CG data are first detected based on clusters of discordant read-pairs, and high confidence SV calls are then characterized whenever local *de novo* assembly for reads in a breakpoint junction is successful (personal communication with Complete Genomics Inc., March 2012).

Various filtering strategies can be used for variants with low quality or with high frequency in a set of controls to further improve specificity of variant calling and delineation of disease-causing changes. Control data that is produced with the same NGS technology and analyzed with the same tools can remove platform and analysis specific sequence artifacts, whereas controls selected from the same population are important in removing common polymorphisms (MacArthur *et al.*, 2014).

3 Aims of the study

The general aim of this thesis was to develop and apply approaches for efficient analysis of large quantities of epidemiological and molecular data. Mastering these approaches allowed characterization of genetic causes associated with various human diseases. The specific aims of studies I-IV were:

I To identify genetic basis of right atrial isomerism in a Finnish family using linkage analysis combined with a candidate-gene approach

II To elucidate molecular basis of a novel intellectual disability syndrome utilizing linkage analysis, homozygosity mapping and whole-genome sequencing

III To analyze whole-genome sequencing and gene expression data to characterize genetic alterations in uterine leiomyomas

IV To develop and utilize novel approaches to systematically analyze the Finnish Cancer Registry data, and to identify potentially novel cancer susceptibility conditions and families

4 Materials and methods

4.1 Study materials

4.1.1 Isomerism family and samples (I)

A Finnish family with exceptionally many offspring with RAI had been previously identified and reported by Eronen *et al*. (2004). Five out of seven siblings were affected with RAI, asplenia and situs anomalies, and had succumbed to complex heart defects before the age of two years. The disease phenotype of the affected children was present already at birth, and the abnormal arrangement of abdominal and thoracic organs, including heart, was very similar in all cases (see Table 1 in the original publication I). The healthy parents (I-1 and I-2, see Figure 1 in the original publication I) were not known to be closely related and no occurrences of laterality defects were seen in the family history. The healthy children (II-5 and II-7) and parents were examined by cardiac and abdominal ultrasound with normal findings.

Genomic DNA was extracted from blood samples of the healthy parents and two healthy children, from fibroblast cell lines of three affected children (II-2, II-4 and II-6) and from paraffin embedded tissue samples of two affected children (II-1 and II-3). Also blood samples from six siblings of the mother and from two sisters of the father were collected and genomic DNA extracted.

DNA samples from 346 anonymous donors of the Finnish Red Cross blood service were used as controls in study I. These were selected from the Jyväskylä (n=125), Lappeenranta (n=82), Kotka (n=74) and Kuopio (n=65) regions. Also 271 control DNA samples from the UK Caucasian blood donors were utilized in the *GDF1* mutation screening (Human Random Control DNA panels, Sigma-Aldrich, Saint Louis, MO, USA). Additional normal tissue or blood samples from 278 CRC patients were chosen for the mutation screening from a sample series collected since 1994 at the Jyväskylä Central Hospital.

4.1.2 Intellectual disability family and samples (II)

Six patients with an unexplained syndromic form of severe ID developed during infancy (see Table 1 in the original publication II) were identified in the same village in the north-east of Finland. The patients were born healthy with normal birth weights from uneventful pregnancies at normal gestational weeks to healthy parents. Etiological investigations including chromosome and copy number analysis and metabolic screening did not reveal the cause of ID in the clinic. Fragile-X DNA analysis of three patients and Prader-Willi methylation analysis of five patients were also negative. Subsequent genealogical work revealed that the patients in the four sibships belong to a large consanguineous family (see Figure 1 in the original publication II). One patient had three deceased siblings with ID, but no patient records were available and no autopsies were performed. Furthermore, one deceased individual (X-1) with severe ID was identified in the extended pedigree, but he did not exhibit the same clinical phenotype as characterized in the other six patients. In addition, we recruited to the study 15 patients living in the north-east of Finland who were diagnosed with severe ID of unknown etiology and had varying degree of compatibility with the patients in the study family.

Individuals and sample materials included in the study are listed in the Supplementary Table 2 in the original publication II. Blood derived RNA was extracted with Paxgene Blood RNA kit (Qiagen, GmbH, Hilden, Germany) and purified with RNeasy MinElute clean up kit (Qiagen). Quality and quantity of DNA and RNA samples were measured before WGS library

preparation and gene expression array hybridization. cDNA was synthesized from RNA samples using random primers (Promega, Madison, WI, USA), M-MLV reverse transcriptase enzyme (Promega) and RNAase inhibitor (Promega).

4.1.3 Leiomyoma samples (III)

A set of 38 uterine leiomyomas and corresponding normal myometrium tissues were collected from 30 patients at hysterectomy at the Helsinki University Central Hospital and snap frozen in liquid nitrogen while fresh. Genomic DNA and total RNA were extracted from the tissue samples with FastDNA Kit (MP Biomedicals LLC, Solon, OH, USA), and TRIzol Reagent (Invitrogen, Carlsbad, CA, USA) or Tri Reagent RT (Molecular Research Center, Inc., Cincinnati, OH, USA), respectively. RNA purification was done with RNeasy MinElute clean up kit (Qiagen). Quality and quantity of the samples were measured before WGS library preparation and gene expression array hybridization.

4.1.4 Patient data in the Finnish Cancer Registry (IV)

The Finnish Cancer Registry (FCR) is a nation-wide database collecting data on incident cancers and cancer deaths in Finland since 1953. Reporting of new cancer incidences was made obligatory for physicians, hospitals, and pathology and hematology laboratories in 1961. FCR has high quality data in terms of completeness and accuracy (Pukkala, 2011; Teppo *et al.*, 1994).

In study IV, we considered 1,175,040 neoplasms registered in FCR until 2011, including 212,685 cases of certain commonly registered precancerous lesions such as basal cell carcinoma of skin and polycythemia vera (registered since 1969). Each record of cancer incidence was classified according to the International Classification of Diseases (ICD) oncology, 3rd edition (ICD-O-3) in the FCR data. We classified tumor types to broader topography and morphology groups before systematic clustering, as described in section 4.5.1. A total of 878,593 patients had personal identity code (PIC), that is, they had not died before 1967, and thus their information on family names at birth and municipalities of birth could be obtained from NPR.

4.2 Array-based methods

4.2.1 SNP array data analysis and genetic mapping (I, II)

In study I, genome-wide SNP genotyping of six individuals (I-1, I-2, II-2, II-4, II-5 and II-6) was performed with Illumina's HumanCNV370 DNA Analysis BeadChip (Illumina Inc.) at the Institute of Molecular Medicine Finland (FIMM) (Helsinki, Finland). Genotype calling was carried out with the BeadStudio software (Illumina Inc.) using the default GenCall score cutoff (0.15) derived from Illumina's GenCall application (v 6.3.0). In study II, genome-wide SNP genotyping was performed with Illumina's HumanOmni2.5 chip (Illumina Inc.) from six patients' DNA (XI-3, XI-7, XII-1, XII-4, XII-5 and XII-6) at Estonian Genome Center (Tarto, Estonia). Genotype calling was carried out with the GenomeStudio software (Illumina Inc.) using the default GenCall score threshold (0.15) and GenCall application (v 6.3.0).

In study I, a multipoint parametric linkage analysis was performed for autosomes with MERLIN (v 1.1.2) (Abecasis *et al.*, 2002) on a CSC server (CSC - IT Center for Science Ltd., Finland). The error detection algorithm included in MERLIN was used to flag unlikely genotypes from the data. The produced error files and the Merlin linkage files were run with pedwipe in MERLIN to erase genotypes flagged as problematic. The parametric linkage analysis using the wiped linkage files was performed under a rare recessive disease model: a disease allele frequency of 0.0005, and penetrances of 0.0, 0.0 and 1.0 for individuals with 0, 1 and 2 copies of the disease allele, respectively. Penetrance was set to 1.0 for two disease

alleles and 0.0 for other options, since the phenotype of the affected children was seen already at birth, and the parents that presumably were carrying one disease allele were confirmed to be healthy. No phenocopies were expected. Genetic map (deCode) distances were derived from Illumina's annotations for the HumanCNV370 panel (Illumina Inc.). SNP allele frequencies in the Finnish population were calculated from control genotypes of 265 healthy individuals provided by the Nordic Centre of Excellence in Disease Genetics consortium.

In study II, a multipoint parametric linkage analysis with MERLIN (v 1.1.2) (Abecasis *et al.*, 2002) was performed with four patients (XI-3, XII-4, XII-5 and XII-6). Problematic genotypes were wiped from the Merlin linkage files as described in study I. Recessive inheritance model with full penetrance and no phenocopies was used for autosomes along markers spaced at 0.25 cM intervals. Marker positions were derived from the HapMap phase II genetic map and SNP allele frequencies from the HapMap phase II data set of the CEPH (Utah residents with ancestry from Northern and Western Europe) population (CEU). Stretches of markers with a LOD score over three were extracted from the results and merged if distance between consecutive stretches was less than 1 cM.

In study II, SimWalk2 (v2.91) (Sobel and Lange, 1996) was used to calculate a parametric LOD score for the disease locus in chromosome 3 with the extended pedigree information. Three candidate disease-causing variants were Sanger sequenced from all the family members from whom DNA was available, and the genotypes of one of the variants were used in the two-point linkage. The disease allele frequency of 0.005 was used based on the allele frequencies in low-coverage sequencing data from the north-east of Finland (402 individuals). Linkage input files for SimWalk2 were generated with Mega2 (Mukhopadhyay *et al.*, 2005).

In studies I and II, linkage-compatible regions were further combined with consecutive homozygous genotypes in affected individuals in order to determine regions of recessive founder mutations.

4.2.2 Gene expression analysis (II, III)

Affymetrix GeneChip Human Exon 1.0 ST microarrays (Affymetrix) were utilized to measure whole-transcriptome expression from peripheral blood of five patients and five healthy siblings in study II, and from 38 leiomyomas and the corresponding myometrium from 30 women in study III. The same leiomyoma and myometrium tissue samples were included in WGS (Table 1).

In studies II and III, signal intensity data from the arrays was first RMA normalized using Brainarray custom CDF files for gene-specific annotation of the probes.

In study II, Student's t-test (two-tailed distribution, two-sample equal variance) and log ratio were calculated for each of the 36,354 genes examined with the array. Pathway enrichment analysis was calculated for the most significant genes with p-value <0.1 (n=831, 2.3% of all genes) with the Ingenuity Pathway Analysis (IPA) software (Build version 220217, Ingenuity Systems, Inc.).

In study III, similarities between the gene expression of tumor samples were assessed using unsupervised hierarchical clustering analysis with the top 1% most variable genes (n=372) which were selected based on the coefficient of variation (ratio of the standard deviation to the mean) calculated across all tumor samples. Statistical testing with three-way analysis of variance (ANOVA) was performed to identify genes that were uniquely expressed in leiomyomas of different genetic background. False discovery rate (FDR) method was used to control p-values for multiple testing (Benjamini and Hochberg, 1995). Pathway enrichment

analysis for *HMGA2*/*HMGA1* overexpressing samples was calculated for the most significant genes using p-value <0.1, and fold change <-1.4 or >1.4 cutoffs (n=440, 1.2% of all genes) with the Ingenuity Pathway Analysis (IPA) software (Build version 302937, Ingenuity Systems, Inc.).

4.3 Fragment analysis

In study I, additional genotyping was carried out using genomic DNA extracted from the paraffin embedded tissue samples of the two affected children (II-1 and II-3), and using the DNA samples of the family members included also in the SNP genotyping. Di-nucleotide microsatellite repeats for four regions with candidate genes were selected from the University of California, Santa Cruz (UCSC) Genome Browser (Human Mar. 2006 Assembly). Lengths of the di-nucleotide repeat units varied from 19 to 26. Primers to the selected microsatellite repeats were designed with Primer3 (v 0.4.0) (Rozen and Skaletsky, 2000). PCR amplification of the samples was done with AmpliTaq Gold DNA polymerase (Applied Biosystems, Foster City, CA, USA). PCR products were prepared for fragment analysis by adding Hi-Di formamide, including GeneScan-500 LIZ size standard (Applied Biosystems). Fragment analysis run was performed at FIMM Genome and Technology Centre as a service. Fragment analysis results were analyzed with GeneMarker software (v 1.4) (SoftGenetics LLC, State College, PA, USA).

4.4 Sequencing methods

4.4.1 Whole-genome sequencing data analysis (II, III)

In studies II and III, genomic DNA libraries were prepared and sequenced according to Illumina (Illumina Inc.) or CG (Complete Genomics Inc.) paired-end sequencing service protocols. The Illumina NGS data consisted of short-insert paired-end reads with 100 bp read length, and with at least 30-fold average sequencing coverage. CG sequencing service was conducted with 40-fold average coverage and with 90% call rate for diploid loci on the human reference genome (Build 37).

WGS was performed on one patient's (XII-1) sample in study II, and on 38 leiomyomas and the corresponding myometrium from 30 women in study III (Table 1). Twelve tumor-normal pairs in study III were sequenced with the Illumina NGS technology, and the rest of the WGS data were produced by CG. All WGS data were analyzed for SVs, CNAs, SNVs and indels. Variant calling was performed only for the Illumina WGS data as described next in section 4.4.1.1. CG provided us with high confidence variant calls that were used as such in the data filtering described in section 4.4.1.2. Variants with high specificity were prioritized in all WGS data analyses. One tumor-normal pair in study III was sequenced with both Illumina and CG sequencing platforms to evaluate whether the variation calls were comparable between the platforms.

Table 1. Uterine leiomyoma and myometrium samples included in the whole-genome sequencing with Illumina and/or Complete Genomics (CG) in study III

Tumor sample	Normal sample	*MED12* mutation (somatic)	*FH* mutation ([a]germline and [b]somatic)	CG	Illumina
M1m3	M1N	c.130G>A, p.G44S	wt	x	
M4m3	M4N	wt	c.738+3A>G[b], and LOH[b]	x	
M5m1	M5N	c.131G>C, p.G44A	wt	x	
M12m1	M12N	c.128A>C, p.Q43P	wt	x	
M17m1	M17N	wt	wt	x	
M18m1	M18N	wt	wt	x	
M29m2	M29N	c.131G>A, p.G44D	wt	x	
M32m1	M32N	wt	c.715G>A, p.A239T[b], and LOH[b]	x	
M32m8	M32N	c.130G>A, p.G44S	wt	x	
M38m5	M38N	wt	wt		x
M44m1	M44N	wt	wt		x
M44m2	M44N	wt	wt		x
M49m1	M49N	c.131G>C, p.G44A	wt	x	
M68m1	M68N	c.131G>A, p.G44D	wt	x	
MY1m1	MY1N	wt	wt	x	
MY9m3	MY9N	c.100-5_c.130del36, loss of splice acceptor	wt	x	
MY10m3	MY10N	wt	wt	x	
MY16m1	MY16N	c.130G>T, p.G44C	wt	x	
MY18m1	MY18N	c.130G>A, p.G44S	wt		x
MY18m2	MY18N	wt	wt		x
MY18m3	MY18N	wt	wt		x
MY22m1	MY22N	wt	wt	x	
MY23m1	MY23N	c.138_152del15, p.N47_V51del	wt		x
MY23m2	MY23N	wt	wt		x
MY23m3	MY23N	c.107T>G, p.L36R	wt		x
MY23m4	MY23N	wt	wt		x
MY24m3	MY24N	wt	wt	x	
MY29m1	MY29N	c.131G>T, p.G44V	wt	x	
MY30m1	MY30N	wt	wt	x	
MY33m2	MY33N	c.100-8T>A, p.E33_D34insPQ	wt	x	
MY45m5	MY45N	c.100-8T>A, p.E33_D34insPQ	wt	x	
MY46m1	MY46N	wt	wt	x	
MY47m1	MY47N	wt	wt	x	
MY48m1	MY48N	wt	wt	x	
MY64m1	MY64N	wt	wt	x	x
MY64m2	MY64N	c.130G>C, p.G44R	wt		x
B7m6	B7N	wt	c.671_672delAG, p.E224fs[a], and LOH[b]	x	
N7m1	N7N	wt	c.1027C>T, p.R343X[a], and LOH[b]	x	

4.4.1.1 Variant calling

SVs were detected with BreakDancerMax (version 1.2) (Chen *et al.*, 2009) from the ELAND alignments provided by the Illumina paired-end sequencing service. Leiomyoma tumor SVs were called with more stringent criteria requiring higher mapping quality score (-q 65) and more discordant reads (-r 4) than in the normal myometrium SV calls (-q 1 -r 1).

The ELAND alignments were realigned with the GATK IndelRealigner using target intervals that were created around dbSNP 135 variations and around initial variant calls produced by the GATK UnifiedGenotyper (DePristo *et al.*, 2011). Duplicate reads were excluded with Picard Tools MarkDuplicates (http://picard.sourceforge.net) from the Illumina WGS data before detection of somatic SNVs, indels and CNAs with VarScan 2 (Koboldt *et al.*, 2012). SNVs and indels were called at sites with a minimum coverage of 10 in both tumor and normal sample. Raw copy number data in 500 bp windows were smoothed and segmented requiring at least 2500 bp regions of equal copy numbers in circular binary segmentation algorithm. Copy number ratios were normalized against the average of all segments of a sample.

4.4.1.2 Data filtering and annotation

In study II, high confidence SV calls, and homozygous single nucleotide and indel variations in protein coding genes were extracted from the CG data within the homozygous regions shared between the six patients. Sequencing data from the north-east of Finland (402 individuals, low-coverage data), from the 1000 Genomes project (phase 1) and from in-house NGS controls (191 Finnish individuals) were used to filter out homozygous variations detected in non-affected individuals. After control filtering, functional effects of the protein coding missense changes were predicted with PolyPhen-2 (v2.2.2r398) (Adzhubei *et al.*, 2010) HumVar-trained models, and variations predicted to be benign were excluded.

In study III, the BreakDancer results were filtered to include tumor SV calls in which either plus or minus strand coverages were at maximum of 60 reads at both breakpoints. This was performed to discard regions with suspiciously high sequencing depth indicative of repetitive and problematic genomic regions. Both Illumina tumor SV and CG somatic high confidence SV calls were filtered against (i) all Illumina normal myometrium SV calls, (ii) all CG normal myometrium SV calls, (iii) DGV variants (Release Date 2012-03-29), (iv) Segmental Dups (UCSC track table) (Bailey *et al.*, 2001), and (v) Hi Seq Depth regions (UCSC track table) (Pickrell *et al.*, 2011). In the filtering steps (i), (ii) and (iii), the type of SV and the locations of the both breakpoints flanked by 1 kb up- and downstream had to match for a tumor SV call to be filtered out. In steps (iv) and (v), a tumor SV call was filtered out, if at least one of the breakpoints was located within a 1000 bp window up- and downstream from repetitive genomic DNA defined by segmental duplications and regions of high sequencing depth. Despite the five systematic filtering steps, Illumina tumor SV calls contained breakpoints within repetitive DNA regions and within regions with discordant reads in the paired myometrium samples which were not called by BreakDancer. Furthermore, 84% percent of Illumina tumor SV calls were visually assessed and excluded as false positives after the systematic filtering. Canonical transcripts in Ensembl version 69 were used to define genes recurrently affected by SV breakpoints within or at maximum 250 kb upstream/downstream of TSS.

In study III, copy number segments with tumor-normal ratio <0.3 were assessed as double deletions, 0.3-0.8 as deletions, 1.2-1.7 as duplications and 1.7-2 as triplications. No high level amplifications were observed. Regions ≤2 kb apart were merged, and regions overlapping gaps in the human genome assembly (UCSC Gap track table) were removed. Segments less

than 100 kb were filtered out. Ensembl version 69 annotations were utilized for mapping genes within CNA segment calls.

In study III, somatic Illumina and CG point mutation and indel calls were filtered against all variant calls from the corresponding myometrium samples (n=30), 93 Finnish individuals from the 1000 Genomes Project (phase 1), rs-coded SNPs from the dbSNP (Build 132), and 157 in-house NGS controls. Depth of coverage at a variant site had to be at least 6, and the percentage of mutated reads had to be at least 20. We chose the minimum variant score thresholds 29 for Illumina and 94 for CG based on previous validation experiments of *MED12* and *FH* mutations, and based on comparison of the common variations in the one sample sequenced with both platforms. Only protein coding variants of the canonical transcripts downloaded from Ensembl version 69 were considered. If a gene displayed a SNV or indel call more than once in our tumor series with scores above the thresholds, the call was subject for verification by Sanger sequencing.

4.4.1.3 Detection of interconnected complex chromosomal rearrangements

Complex chromosomal rearrangements (CCRs) were assessed using a computational method utilizing somatic SV and CNA calls. For each tumor, an event graph with nodes and edges was constructed as described in Figure 3 (see also Figure 1A and 1B in the original publication III).

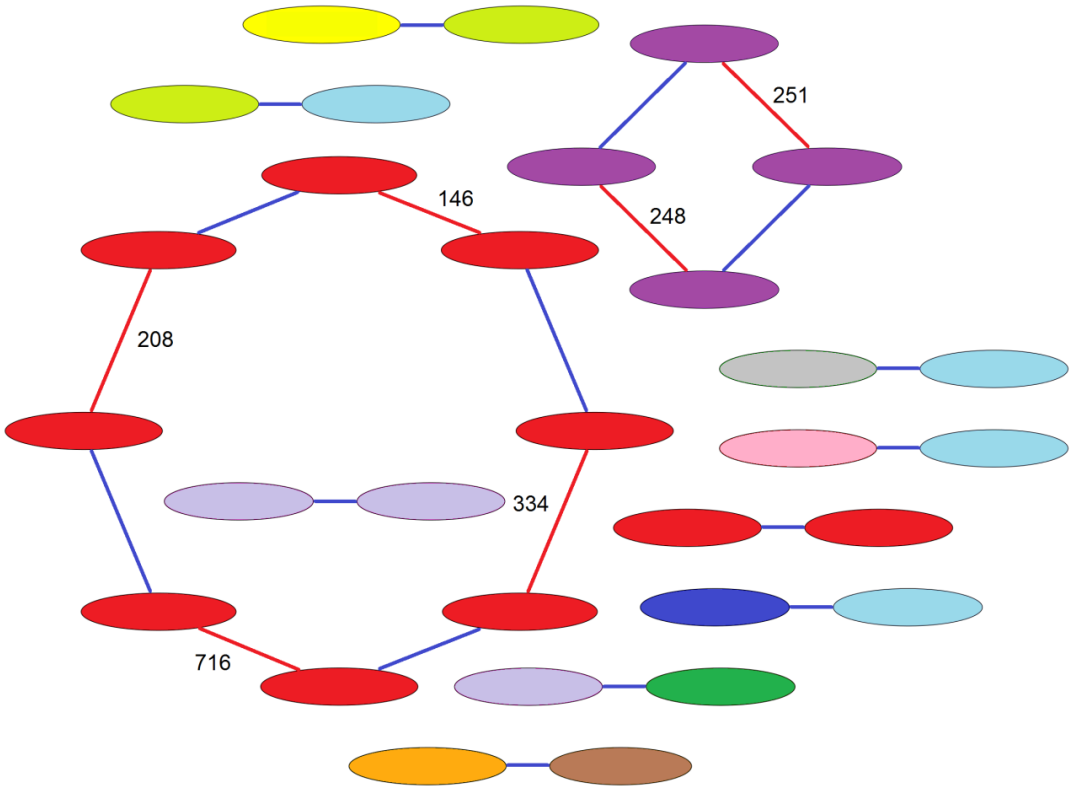

Figure 3. Event graph of somatic structural variations (SVs) in tumor M29m2. Nodes in the graph represent DNA double-strand break ends either left (head) or right (tail) of a breakpoint. Colors of the nodes represent different chromosomes. Nodes are connected by blue edges (i) if break ends are connected to each other by a SV call of discordant read-pairs, or by red edges (ii) if break ends in head and tail orientation are at most 1000 bases apart (numbers refer to the distance between heads and tails). The distance between a head and a tail was decreased by the length of deletion calls if found to be located between heads and tails. In this tumor, one component consisting of eight connected nodes created through four DNA double-strand breaks was identified in chromosome 4 (red nodes) and called as an interconnected complex chromosomal rearrangements event.

Components consisting of connected nodes were identified from each graph. After manual assessment of components, a CCR event was called only for samples with a minimum of six break ends that were interconnected.

4.4.1.4 Assessment of clonally related leiomyomas

Based on the mutation status of *MED12*, we selected to study additional tumors from four individuals (M29, M38, M68, and MY18) which were not included in WGS but were potentially clonally related. Each of these individuals had at least one lesion included in WGS from which unique somatic markers were selected for Sanger sequencing. Altogether 19 leiomyomas were examined: 5 leiomyomas with wild type *MED12* from patient MY18, 4 leiomyomas with wild type *MED12* from patient M29, 2 leiomyomas with wild type *MED12* from patient M38, and 8 leiomyomas with the same *MED12* mutation (c.131G>A, p.G44D) from patient M68.

4.4.2 PCR and Sanger sequencing (I, II, III)

PCR primers were designed with web-based tools (Rozen and Skaletsky, 2000). In study I, the entire coding region or the two mutation sites of the candidate gene, *growth/differentiation factor 1* (*GDF1*), were Sanger sequenced from the blood samples or from the paraffin embedded tissue samples, respectively. In study II and III, candidate SNVs and indel calls detected in the WGS data were verified by Sanger sequencing. In study II, cDNA samples of patients and controls were Sanger sequenced to detect splicing defects in two genes with candidate mutations. In study III, sample MY64m1 was in WGS with both Illumina and CG sequencing platforms, and all 32 SV calls found by either or both of the platforms were validated by Sanger sequencing. To assess the platform specific true positive rate, final SV calls from Illumina samples MY18m2 and MY18m3, and CG sample MY47m1 were also subjects for validation.

PCR amplification was performed using AmpliTaq Gold DNA polymerase (Applied Biosystems) or Phusion DNA polymerase (Finnzymes, Finland). In study I, a modified reaction mixture from Expand Long Template PCR system (Roche Diagnostics, Germany) with GC-Melt Reagent (Clontech, Palo Alto, CA, USA) was used according to Joensuu *et al.* (Joensuu *et al.*, 2007) for fragments with high guanine-cytosine content. The PCR products were purified using the ExoSAP-IT PCR purification reagent (USB Corporation, Cleveland, OH, USA), and the sequencing reactions were performed utilizing the Big Dye Terminator v.3.1 kit (Applied Biosystems). Electrophoresis was performed on a 3730xl DNA Analyzer (Applied Biosystems) at FIMM Genome and Technology Centre (Helsinki, Finland). The sequence graphs were analyzed manually or with the Mutation Surveyor software (versions 3.24 and 3.30, Softgenetics, State College, PA, USA).

4.5 Registry-based data analysis

4.5.1 Systematic clustering of patients (IV)

For each tumor type, systematic clustering was performed based on family name at birth (denoted as N-clusters), and combination of both municipality of birth and family name at birth (MN-clusters). The observed number of patients (O) with each tumor type in FCR was calculated in a stratum defined by municipality (M) (in case of MN-clusters only), family name (N), sex and year of birth. To estimate the expected number of cases (E) in each stratum, gender and birth-year specific number of cases of the tumor type in question were calculated from the NPR-linked subset of FCR. This number of cases was multiplied with the proportion of persons in each stratum in the entire population of the same gender and year of birth calculated from the NPR database. The stratum-specific observed and expected numbers were

added up over the gender and birth-year categories. O/E ratios were calculated for each N and MN category, and their 95% confidence intervals were defined assuming a Poisson distribution of observed numbers.

4.5.2 Estimating familiality with cluster score (IV)

To estimate the familial occurrence of various tumor types, a cluster score was calculated for MN-clusters, which were more likely representing true families than N-clusters. For each tumor type *i*, we selected clusters with the lower limit of confidence interval (CI-low) greater than or equal to α and calculated the number of patients ($X_{i\alpha}$) in these clusters.

The cluster score of tumor type *i* was defined as $X_{i\alpha}$ divided by the total number of patients with the same tumor type (Y_i) per 100,000 persons in Finland. To estimate whether $X_{i\alpha}/Y_i$ was different than the expected ratio Y_iX_α/Y, where X_α/Y is the respective ratio over all tumor types in Finland ($X_\alpha/Y=0.0029$ for $\alpha=10$), two-sided Poisson test was calculated with 95% confidence level. The p-values were adjusted for multiple test comparisons with FDR method (Benjamini and Hochberg, 1995).

4.6 Ethical issues

Studies I and II were approved by the appropriate ethic review committees (Dnro 310/E7/2000, Dnro 55/07/2000 and No. 110/2010), the respective clinical researchers as principal investigators. Written informed consent was received from the patients or from the patients' parents or guardians, and from the healthy individuals in studies I and II. Three series of uterine leiomyoma samples were utilized in study III. The first series (denoted as "M" patients) consisted of samples from anonymous patients included in the study with the approval of the director of the appropriate health care unit. The second series (denoted as "MY" patients) and third series (denoted as "B" and "N" patients with the HLRCC syndrome) consisted of samples from patients with an acquired informed consent.

Tumor patient data and sample collection procedures were evaluated and approved by the Ministry of Social Affairs and Health in Finland (Dnro 53/07/2000) and subsequently by the National Institute for Health and Welfare (THL/1071/5.05.00/2011).

5 Results

5.1 Identification of *GDF1* mutations in right atrial isomerism (I)

Genome-wide SNP genotyping and linkage analysis of the parents, one healthy child and three affected children revealed 15 genomic regions compatible with recessive inheritance (Table 2). All of the fifteen loci had the maximum LOD score of 1.33. The longest homozygosity within the linkage-compatible regions was on chromosome 7 overlapping the centromere. The physical and genetic length of the homozygous region was 9.3 Mb and 1 cM, respectively, reflecting the low recombination rate of the centromeric region. Other homozygous regions did not exceed 1 cM; thus parents were likely carrying different founder mutations. Candidate-gene approach was chosen for prioritization of the regions, and four of the linkage-compatible regions included genes involved in the left-right axis development in model organisms (Table 2) (Levin, 2005; Okumura *et al.*, 2008; Shen, 2007).

Table 2. Linkage-compatible regions and positional candidate genes in study I

Chr	Start position (Mb)[a]	End position (Mb)[a]	Region length (Mb)	Region length (cM)	Candidate genes in model organisms	HGNC Symbol
1	68.46	94.20	25.7	21.96	*Pcl-2* in chick	*MTF2*
3	0.07	5.81	5.7	16.65		
4	59.49	86.38	26.9	18.94		
5	166.88	168.25	1.4	2.84		
6	56.40	85.21	28.8	12.79		
7	33.67	45.53	11.9	13.83	*Myosin31DF* in *D. melanogaster*	*MYO1G*
7	57.40	88.48	31.1	21.76	*Hgf* in chick	*HGF*
8	6.40	10.53	4.1	7.80		
9	33.62	38.77	5.1	5.16		
9	76.32	92.29	16.0	23.82		
14	78.06	78.08	0.02	0.02		
17	77.29	81.04	3.7	12.71		
19	13.33	23.48	10.1	14.45	*Gdf-1/Vg-1* in mouse/frog, *Notch* in mouse/zebrafish	*GDF1*, *NOTCH3*
19	55.97	59.08	3.1	5.61		
22	17.08	19.03	2.0	6.87		

[a] Genomic coordinates were converted from Build 36 to Build 37 with liftOver tool.

The fragment analysis of di-nucleotide microsatellite repeats from additional two affected individuals excluded three of the four candidate loci. All five affected children were shown to share the same alleles within the linkage-compatible region on chromosome 19 (Figure 4). This region contained two candidate genes: *NOTCH3* and *GDF1*. Autosomal dominant *NOTCH3* mutations had been identified in patients with cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (MIM #125310). *GDF1* had

previously been reported to be involved in the left-right patterning in mice with targeted homozygous deletion of the region encoding the mature Gdf1 (Rankin *et al.*, 2000).

Sanger sequencing of the protein coding regions of *GDF1* (NT_011295.11) revealed two truncating mutations that segregated with the RAI phenotype in an autosomal recessive manner (Figure 4). The first mutation resided in exon 8 and caused a premature stop codon (c.681C>A, p.C227X). The second mutation was an insertion of cytosine (c.909insC) in exon 8 leading to a frameshift and truncation (after 22 residues) (see Figure 2 and 3 in the original publication I). The father and the mother were healthy heterozygous carriers of the nonsense and the frameshift mutation, respectively. Affected children (II-1, II-2, II-4, and II-6) were compound heterozygotes for the two mutations in sequencing. The fifth affected child (II-3) was also carrying the mutations based on the microsatellite haplotypes, although DNA extracted from the paraffin embedded specimen was not of sufficient quality for successful sequencing. The healthy children displayed no mutations in *GDF1*.

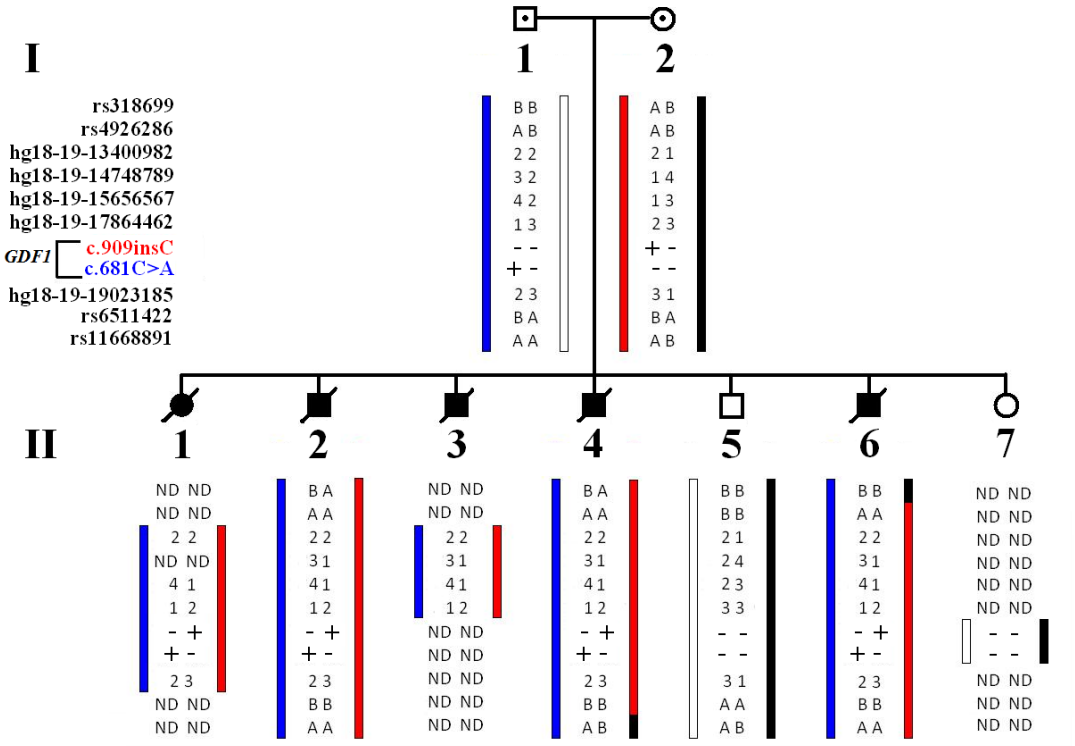

Figure 4. Family pedigree and haplotype analysis from the linkage-compatible region on chromosome 19. The haplotypes associated with the nonsense mutation and with the insertion in *GDF1* are colored blue and red, respectively. Plus denotes mutation and minus wild type allele on the mutation sites. Selected SNP genotypes are derived from Illumina's genotyping panel. Microsatellites are named according to the chromosome number and physical positions in the human reference genome Build 36 (hg18), and scored from 1 to 4 based on the ordered lengths of the PCR fragments. Squares denote men and circles women, solid symbols affected children, symbols with a dot healthy mutation carriers, and open symbols unaffected children. ND: not determined.

The mutation sites identified in the family were screened in anonymous blood donors from the Finnish and UK populations. Among 346 Finnish blood donors, we found two heterozygous carriers of the insertion (c.909insC) and one heterozygous carrier of the nonsense mutation (c.681C>A, p.C227X). The whole coding region of *GDF1* was sequenced from these individuals and no other variations were characterized. Among 271 control

samples from the UK population, one carrier of the nonsense mutation was found confirming the prevalence of the mutation outside Finland.

In order to study the phenotypic effects of heterozygous *GDF1* mutations further, healthy siblings of the parents and Finnish CRC patients were examined for the two mutations by Sanger sequencing. Three siblings of the mother carried the insertion without any indication of heart disease when their health status was interviewed. Among 278 CRC patients, one carrier of the insertion was found, and one patient had a heterozygous 11 bp deletion in *GDF1* (c.793_803delGGCGCTTGTCG). Although GDF1 is a player in the transforming growth factor beta (TGFβ) signaling pathway and binds to Activin receptors (Cheng *et al.*, 2003), truncating mutations in *GDF1* do not seem to predispose to CRC (Fisher´s exact test, two-tailed p-value=1). The clinical records of the two CRC patients heterozygous for the frameshift mutations in *GDF1* revealed normal cardiac development as evaluated by cardiac ultrasound in one and coronary bypass operation-related examinations in the other.

5.2 Genetic mapping of severe intellectual disability syndrome (II)

Six ID patients in four separate core families were identified by clinicians L. Pajunen and E. Rahikkala. The patients had developed a highly similar syndromic form of severe ID of unknown etiology, described in detail in the patient descriptions of the original publication II. All eight healthy parents of the patients were shown to have a common ancestor that lived in the 17th century (see Figure 1 in the original publication II), suggestive of recessive inheritance of a novel ID syndrome. Four patients from two core families were utilized in the initial genome-wide linkage analysis. The longest recessively inherited region was on chromosome 3 encompassing 8.5 cM with the overall LOD score >3 (Figure 5). The second longest linked region was 3.5 cM and located also on chromosome 3. Homozygosity mapping of the SNP genotyping data of all six patients revealed an 11.5 Mb (chr3:42417576-53886650, 5.2 cM) stretch of homozygosity within the longest linked region at 3p22.1-3p21.1 (see Figure 3A in the original publication II).

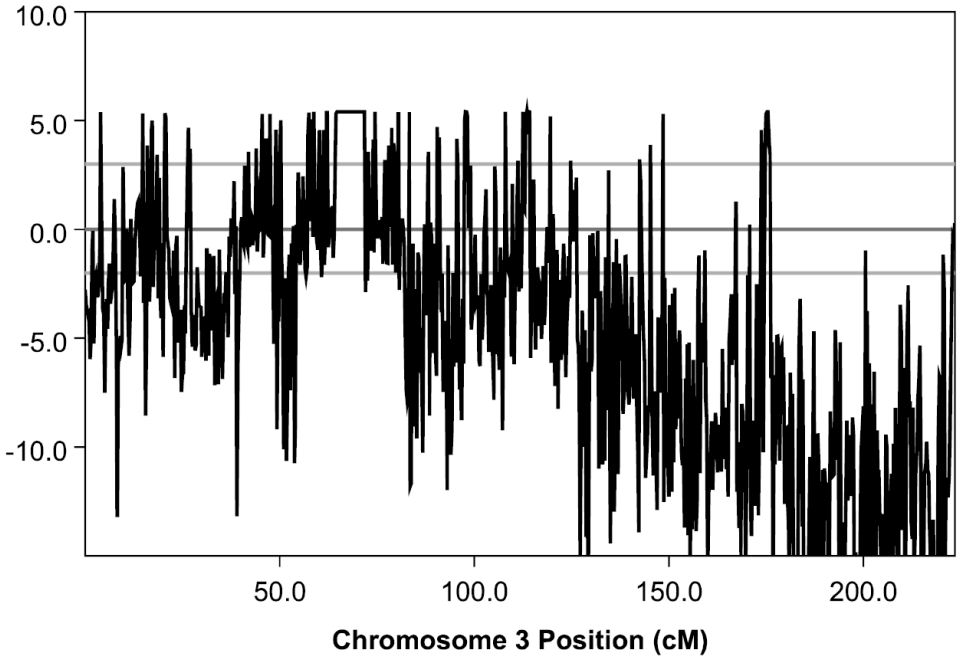

Figure 5. Chromosome 3 displayed the longest region compatible with recessive inheritance of a novel intellectual disability syndrome at positions 64.3-72.8 cM. Y-axis represents the logarithm of odds (LOD) score calculated from the SNP genotyping data of four patients by the MERLIN parametric linkage analysis program.

No SVs were found at 3p22.1-3p21.1 in the WGS data of one of the patients. However, the data revealed three rare homozygous SNVs or indels within the locus that were *in silico* predicted to be probably protein damaging or causing errors in the splicing of protein coding genes (Table 3). All homozygous variants were shared by the six patients as validated by Sanger sequencing. The parents were heterozygous carriers, and the healthy siblings were either heterozygous or wild type. The Sanger sequencing results were utilized in the single point parametric linkage by SimWalk2. The information from the extended pedigree resulted in the final LOD score of 11 for the homozygous locus at 3p22.1-3p21.1.

Table 3. Homozygous variations found at 3p22.1-3p21.1 in whole-genome sequencing after filtering and validated in all six patients by Sanger sequencing in study II

HGNC symbol	Variation[a]	Effect	UniProtKB accession	Protein change	PolyPhen-2	I	II	III
P4HTM	3:49042479 G>A	splice site defect	Q9NXG6-1	R296S+ 297_358del	-	5	1	-
		missense	Q9NXG6-3	R358Q	probably damaging			
USP4	3:49323756_ 49323758 delCTC	splice site defect	Q13107-1	G658del, G658_E659del	-	6	-	-
		splice site defect	Q13107-2	G611del, G611_E612del	-			
TKT	3:53269087 T>C	missense	P29401-2	I189V	probably damaging	3	7	1
		missense	P29401-1	I181V	benign			

[a] Genomic coordinates from Build 37

I No. of heterozygous carriers in the regional controls (402 individuals)

II No. of heterozygous carriers in the 1000 Genomes project data (1092 individuals)

III No. of heterozygous carriers in the in-house next-generation sequencing controls of Finnish origin (191 individuals)

An in-frame loss of exon 6 (c.888_1073del186, p.Arg296Ser+Val297_Arg358del) of the *prolyl 4-hydroxylase transmembrane* (*P4HTM*) transcript (NM_177939.2) was detected in the patients' blood derived cDNA (see Figure 3B in the original publication II). The effect was compatible with the defective 5' splice site recognition and exon definition (Berget, 1995). The other known transcript of *P4HTM* with the presumed p.Arg358Gln change could not been detected in the blood. The homozygous deletion at a splice site of the *ubiquitin specific peptidase 4* (*USP4*) transcripts (NM_003363.3 and NM_199443.2) produced two new transcripts through alternative 3' splice site recognition: one with a 3 bp loss, c.1973_1975delGAG (p.Gly658del), and another with a 6 bp loss, c.1973_1978delGAGAAG (p.Gly658_Glu659del) (see Figure 3C in the original publication II). The third segregating homozygous variation affecting transketolase (TKT) caused protein changes p.Ile181Val (isoform 1, NP_001055.1) and p.Ile189Val (isoform 2, NP_001244957.1) which were benign and probably damaging, respectively. The isoform 2 of TKT has an alternative sequence of 8 amino acids at the position 146 of the isoform 1, but how this alternative sequence affects the prediction remains ambiguous.

The variants displayed 0.3-0.7% allele frequencies in the low-coverage WGS data derived from 402 individuals from the same region as the patients (Table 3). No homozygotes were found among any of the controls. We also studied the three candidate variants in additional 15 patients from the north-east of Finland, and one deceased ID patient identified in the extended pedigree with severe ID of unknown etiology. All 16 patients presented with varying degree of compatibility with the novel syndrome. None of these patients carried the candidate sequence changes.

An enzyme assay showed normal TKT activity in the patients' lymphoblasts as compared to non-affected controls. Elevated levels of sugars and polyols were noted in one patient's urine and plasma as described earlier in ribose-5-phosphate isomerase (RPI) deficiency (Huck *et al.*, 2004). However, other patients had no change in sugar and polyol levels. Immunoblotting and real-time PCR experiments were performed to study the abundance of P4H-TM and hypoxia-inducible factor-1 alpha (HIF-1α) and the expression of HIF-1α target genes in normoxic or hypoxic conditions in the patients' lymphoblasts as compared to healthy siblings. The *P4HTM* transcript with the loss of exon 6 showed reduced induction in hypoxia, but the P4H-TM protein levels were unchanged in the patients' lymphoblasts. The HIF-1α protein levels and expression of the HIF-1α target genes were also unchanged (see Supplementary Figure 2 in the original publication II). The experiments were done in collaboration with researchers from University of Oulu (Finland) and VU University Medical Center (Amsterdam, the Netherlands) and are described in detail in the original publication II.

In gene expression analysis from the blood, the two most significantly altered pathways in five patients as compared to five healthy siblings were relaxin signaling and CREB [cyclic adenosine monophosphate (cAMP) response element-binding protein] signaling in neurons. Also cAMP-mediated signaling and G-protein coupled receptor signaling were among the five most significantly altered pathways. USP4 has been shown to regulate a G-protein coupled receptor at the cell surface (Milojevic *et al.*, 2006), which could be an upstream event leading to the altered cAMP-mediated and CREB signaling. However, we could not detect changes in cAMP concentrations with immunoassay kit, or in the protein levels of phosphorylated CREB and total CREB with immunoblotting (see Supplementary Figure 3 in the original publication II). The immunoassay and immunoblotting experiments were performed as described earlier (Tuominen *et al.*, 2014) using protein extracts from lymphoblastoid cell lines of patients as compared to healthy non-carrier siblings.

5.3 Molecular genetic characteristics of uterine leiomyomas (III)

5.3.1 Landscape of somatic alterations and complex chromosomal rearrangements

The previously validated *MED12* and *FH* mutations (Table 1 in section 4.4.1) were the only somatic SNVs or indels recurrently affecting protein coding genes in the set of 38 leiomyomas included in WGS. The simple chromosomal rearrangements that were described previously in the literature were detected also in our WGS data. The large-scale deletions causing LOH on chromosome band 1q43 at the *FH* locus were detected in four tumors with germline (N7m1 and B7m6) or somatic (M32m1 and M4m3) *FH* mutation (Table 1 in section 4.4.1). Three tumors displayed the balanced t(12;14)(q15;q24) translocation and four tumors a large deletion on chromosome 7q-arm.

The WGS data revealed abundant complex chromosomal rearrangements in uterine leiomyomas. Rearrangement breakpoints were located near breakpoints from other rearrangements or deletions assessed as chained event graphs (see Figure 1A and 1B in the original publication III). These interconnected CCRs were detected most frequently in tumors lacking *MED12* and *FH* mutations (12/16), in three *MED12*-mutant tumors (3/16) and in none of the *FH*-deficient tumors (*MED12* and *FH* mutant tumors vs. other tumors, Fisher´s exact test, two-tailed p-value <0.001). Five tumors with CCR events (MY10m3, MY23m4, MY46m1, MY47m1 and MY64m1) displayed 20 or more breakpoints in a single chromosome; thus they were typical examples of the recently described chromothripsis phenomenon (Stephens *et al.*, 2011).

Sanger sequencing of all SVs from three leiomyomas with a high number of breakpoints verified CCR events. Validation showed 93% (64/69) and 98% (99/101) true positive rate for SVs detected in the Illumina and CG data, respectively. To compare data quality between the two different platforms and to assess false negative rate, MY64m1 was sequenced using both Illumina and CG. Altogether 32 SV calls were found with both platforms and validated. Only eight and five validated SV calls were found to be uniquely called in the Illumina and CG data, respectively.

CCR events in two tumors had created rearrangements between chromosome 12 and 14 combining the 5´ end of the *RAD51B* gene with full length *HMGA2*, a phenomenon which is detected likewise in the balanced t(12;14)(q15;q24) translocations. Three other tumors displayed also CCR events resulting in rearrangements at *HMGA2* or *HMGA1* loci, but without *RAD51B* involvement. One of these rearrangements removed the 3' UTR of *HMGA2*, which is the target sequence for the microRNA repressor let-7b, providing a mechanism for upregulation of *HMGA2* (Mayr *et al.*, 2007). The oncogenic upregulation was detected in all tumors with *HMGA2*/*HMGA1* rearrangements in the gene expression data. The rearrangements lacking *RAD51B* involvement caused the lowest *HMGA2* upregulation, suggesting that *RAD51B* is accompanied by an effective enhancer for *HMGA2*. Based on deletion and breakpoint mapping at the *RAD51B* locus, we pinpointed the minimal region containing a possible enhancer for *HMGA2* to chr14:68217257-68760115 (see Supplementary Figure 6 in the original publication III).

Recurrent chromosomal changes were detected also at the *COL4A5*/*COL4A6* locus on chromosome Xq22 in tumors without *MED12*, *FH* or *HMGA2*/*HMGA1* alterations. This locus is known to be disrupted in persons with Alport syndrome and diffuse leiomyomatosis (ATS-DL, MIM #308940). A characteristic simple deletion reported previously in ATS-DL patients was detected in one uterine leiomyoma (see Figure 3A in the original publication III). In two tumors, CCRs resulted in the fusion of the 3´ends of the two collagen genes (see Figure 3B and C in the original publication III). These three tumors showed on average 5.6-fold upregulation of the *IRS4* gene which is located adjacent to *COL4A5*.

In tumor MY47m1, the CCR event disrupting *COL4A5/COL4A6* was detected only in CG somatic low confidence SV calls, and confirmed by Sanger sequencing. This tumor had the highest number of breakpoints and chromothripsis (Figure 6A). One of the breakpoints in chromosome 11 located 215 kb upstream of the cell cycle progression gene *CCND1* (encoding cyclin D1) that was upregulated by 16-fold as compared to the corresponding myometrium. Because loss of p53 was earlier associated with increased frequency of chromothripsis (Rausch *et al.*, 2012), we studied carefully all the breakpoints that would disrupt *TP53*. One tumor (MY10m3) had a complex rearrangement disrupting one allele of *TP53* and a chromothripsis-like remodeling event on chromosome 17. The SV hit both *TP53* and *NF1* genes at the same time as validated by Sanger sequencing.

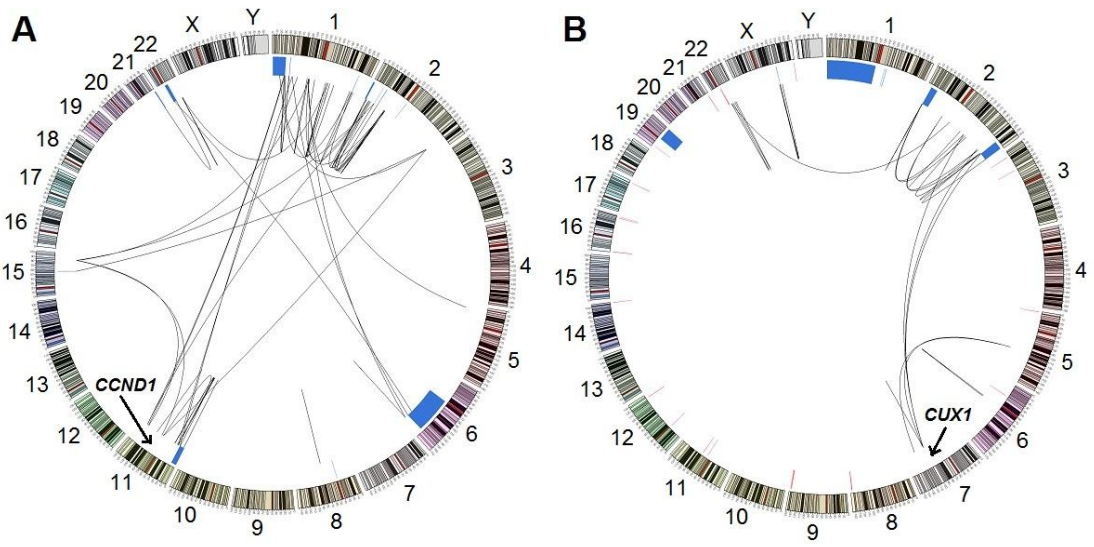

Figure 6. Circos plots of somatic structural variations and copy number alterations detected in samples MY47m1 (A) and MY23m4 (B) with complex chromosomal rearrangements resembling chromothripsis. Chromosomes are organized circularly in the outer ring. Black lines in the circos plots indicate structural variations, and blue segments deletions.

Some tumors contained more than one independent CCR event as shown by spatial or temporal difference of interconnected rearrangements. Two spatially separate CCR events were present in MY18m2 with equal discordant read frequency (see Figure 1C in the original publication III). The other event involved chromosomes 1, 2 and 20, and the other event involved chromosomes 12 and 14 creating the translocation between the *HMGA2* and *RAD51B* loci. The same tumor also displayed a temporally separate CCR event involving chromosome 5 (see Figure 1D in the original publication III). Another tumor, MY23m4 displayed two independent CCR events affecting both copies of chromosome 7 and the *cut-like homeobox 1* (*CUX1*) gene. The interconnected rearrangements in MY23m4 were affecting chromosome 2 and chromosome 7 in one event, and chromosome 5 and another copy of chromosome 7 in the other event (Figure 6B).

The most recurrently affected gene by deletions and breakpoints was *CUX1* on chromosome 7q-arm (see Supplementary Figure 7 in the original publication III). The tumor MY23m4 with two independent CCR events, each disrupting one copy of chromosome 7, had the lowest expression of *CUX1* among all tumors (see Supplementary Figure 5 in the original publication III). The tumors with *CUX1* changes did not cluster together in hierarchical clustering analysis of gene expression data. Instead, the global gene expression profiling revealed three separate leiomyoma subgroups: *MED12* mutation-positive, *FH*-deficient and *HMGA2/HMGA1*-overexpressing samples, which clustered in distinct branches (see Figure 2 in the original publication III) as previously suggested in separate studies (Hodge *et al.*, 2012; Mäkinen *et al.*, 2011b; Vanharanta *et al.*, 2006). Interestingly, the most up-regulated gene in *MED12* mutation-positive tumors was *RAD51B*. Furthermore, tumors with *HMGA2/HMGA1* overexpression displayed upregulation of genes involved in the cell cycle G1/S checkpoint regulation, such as *CCND1*, *CCND2*, *CCND3* and *CDK6*.

5.3.2 Clonal origin of multiple tumors

Several identical chromosomal changes were detected in the WGS data of two tumor pairs from two patients (MY18 and M44). The other tumor pair, MY18m2 and MY18m3, shared over 20 identical rearrangements which were not detected in the third tumor (MY18m1) that

was also included in the WGS experiment from the same patient. The other tumor pair, M44m1 and M44m2, shared nine identical rearrangements. MY18m2 and M44m1 had fewer SV calls detected on chromosome 5 and chromosome 7, respectively, as compared to their clonally related tumors MY18m3 and M44m2. However, these rearrangements were present in the aligned WGS reads of MY18m2 and M44m1 with lower discordant read frequency than the threshold used in the variant calling. In addition to chromosome 5 changes, MY18m3 had CNA calls on chromosomes 4 and 19 not called in MY18m2. Since chromosome 12 and 14 changes as well as chromosome 1, 2 and 20 changes were detected in both MY18m2 and MY18m3 at similar levels, we considered the changes in chromosome 4, 5 and 19 to be only subclonally present in MY18m2, as detected by the calculation of average log2 copy number ratios (tumor versus normal) for the deleted regions (Table 5). This pattern of changes suggested MY18m2 to be a primary tumor from which a cell lineage with the subclonal somatic changes in chromosome 4, 5 and 19 had formed the secondary tumor MY18m3.

Table 5. Average log2 copy number ratios for the deleted regions in the clonally related tumors MY18m2 and MY18m3

Chromosome 4, a deletion on p-arm, subclonal in MY18m2, chr4: start-49,660,117	
Sample	**Average log2 ratio**
MY18m2	-0.141
MY18m3	-0.577
Chromosome 19, a deletion on q-arm, subclonal in MY18m2, chr19: 27,681,783-end	
Sample	**Average log2 ratio**
MY18m2	-0.117
MY18m3	-0.484
Chromosome 5, CCR related deletions, subclonal in MY18m2, chr5: 107,392,713-149,571,004 and chr5: 149,752,077-158,199,892	
Sample	**Average log2 ratio**
MY18m2	-0.162
MY18m3	-0.715
Chromosome 1, a CCR related deletion, chr1: 882,685-29,447,281	
Sample	**Average log2 ratio**
MY18m2	-0.718
MY18m3	-0.722
Chromosome 14, CCR related deletions, chr14: 39,424,193-43,740,449 and chr14: 63,744,555-68,216,937	
Sample	**Average log2 ratio**
MY18m2	-0.715
MY18m3	-0.795

To find additional tumors that were clonally related, Sanger sequencing of somatic point mutations was performed from tumors selected based on concordant *MED12* mutation status. Three additional tumors (MY18m4, MY18m5 and MY18m6) from patient MY18 were shown to be clonally related to MY18m2 and MY18m3. Patient M38 had one additional tumor clonally related to the tumor M38m5. We did not find additional tumors of patients M29 and M68 to be clonally related to the tumors M29m2 and M68m1, which were included in WGS.

Altogether two tumors from patients M44 and M38, and five out of six tumors from patient MY18 were shown to be clonally related.

5.4 Familial aggregation of tumor types in Finland (IV)

Two systematic clustering efforts for 878,593 cases in the FCR database were performed: one based on tumor type, municipality of birth and family name at birth (MN-clusters), and another on tumor type and family name at birth (N-clusters). As a result, we identified 25,910 MN-clusters and 12,695 N-clusters representing 183 tumor types (Figure 7).

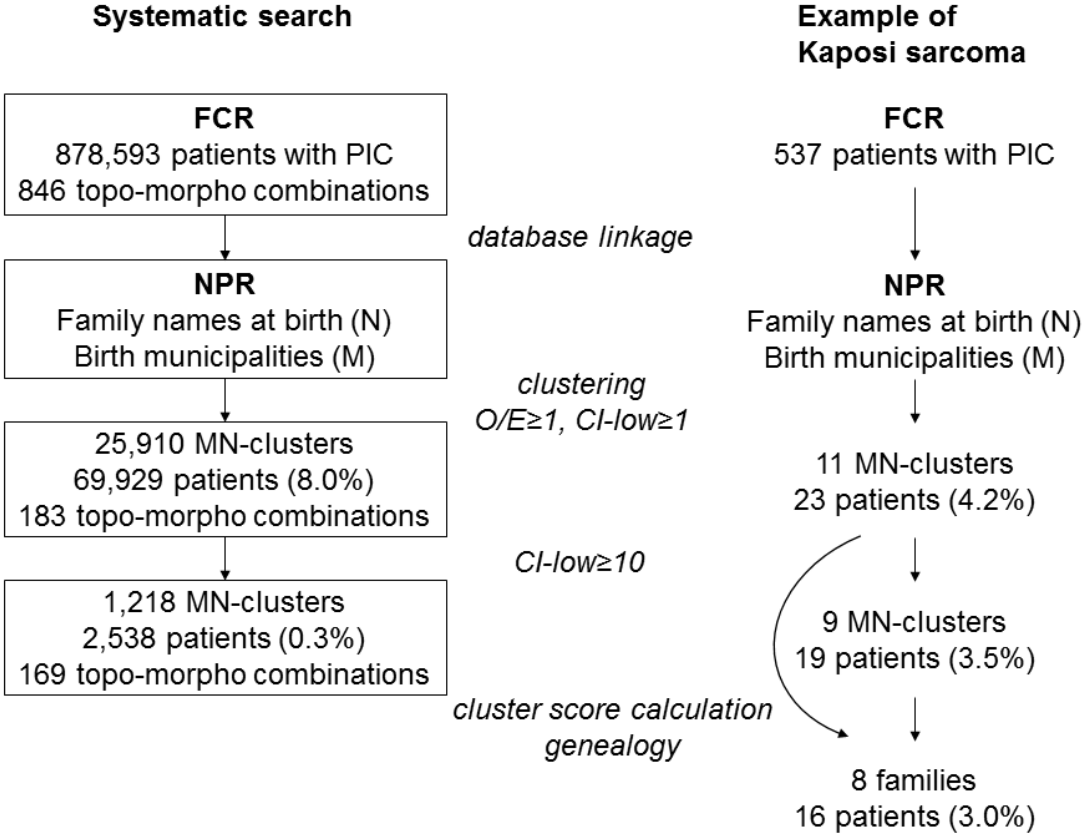

Figure 7. Procedure of the systematic clustering with the example of Kaposi sarcoma. Tumor types in the Finnish Cancer Registry (FCR) were defined by unique combinations of topography and morphology (topo-morpho) classifications. Patients with personal identity code (PIC) were linked to the National Population Registry (NPR). Eight percent of the patients clustered based on tumor type, municipality of birth and family name at birth (MN-clusters). MN-clusters with the strongest confidence, with the lower confidence limit (CI-low) greater than or equal to ten, were included in the cluster score calculation. Kaposi sarcoma patients clustered frequently in MN-clusters with CI-low≥10 (3.5%) as compared to all patients in FCR (0.3%) and, thus, were subject for thorough genealogy work, which revealed eight families with two first-degree relatives diagnosed with Kaposi sarcoma.

We considered only the most significant MN-clusters in the cluster score calculation that demonstrated our familial aggregation measure for the tumor types. Twenty tumor types, ranked according to cluster score, displayed FDR adjusted p-values less than 0.0001 (see Table 1 in the original publication IV). Among the top five tumor types showing strongest evidence for familial clustering were four phenotypes that have predisposition genes identified earlier. These included hemangioblastoma (cluster score 4.98), medullary thyroid carcinoma (cluster score 3.55), pancreatic neuroendocrine tumors (cluster score 1.40), and nephroblastoma (cluster score 1.26), which were indicative of patients with the Von Hippel-Lindau syndrome (VHL, MIM #193300), Multiple Endocrine Neoplasia type 2 (MEN2A,

MIM #171400, and MEN2B, MIM #162300), Multiple Endocrine Neoplasia type 1 (MEN1, MIM #131100) and Wilms tumor 1 (WT1, MIM #194070), respectively.

Kaposi sarcoma was the highest scoring tumor type for which prevalent predisposition gene mutations were not previously known. Relatives of the KS patients in the MN-clusters were traced back at least three generations to confirm kinship of the clustered cases. The genealogy revealed that 16 patients out of 23 (70%) were first-degree relatives within the MN-cluster, and most of these were found in clusters with CI-low≥10 (Figure 8). One of the families displayed five KS patients in two generations (see Figure 2 in the original publication IV). Three affected siblings in the family had the same family name at birth. One of them was born in a different municipality and, therefore, could be linked to the family through the respective N-cluster (Figure 8). An affected cousin was connected to the family through genealogy work, and the mother's KS diagnosis was confirmed from radiotherapy records after it was reported by her daughter.

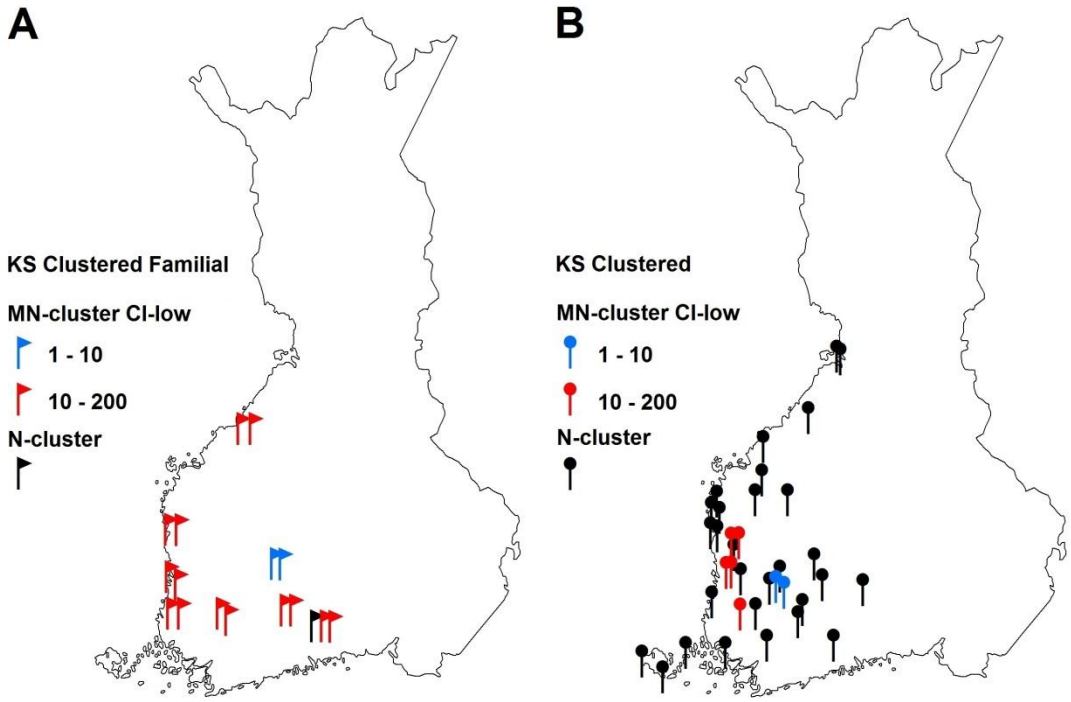

Figure 8. Geographical distribution of the birth places of Kaposi sarcoma (KS) patients in clusters. Clustering was performed based on municipality of birth and family name at birth (MN-clusters), or based on family name at birth only (N-clusters). (A) Sixteen familial cases were identified through genealogy of MN-clusters. One case was connected to a family of two cases based on N-clustering. (B) Familiality could not be confirmed for 36 cases in MN- or N-clusters. Most of the KS patients in the clusters were born in the western part of Finland. Flags and pins are colored blue and red according to the lower confidence limit (CI-low) of the respective MN-clusters.

The overall KS incidence was studied utilizing the small-area mapping method developed in FCR (http://astra.cancer.fi/cancermaps/suomi5308/). Patients were shown to accumulate to the Western Finland (see Figure 3 in the original publication IV), as also illustrated by the geographical distribution of the clustered KS cases (Figure 8).

6 Discussion

Thorough clinical characterization of patients and good sample materials enabled successful research on various phenotypes conducted in this thesis (Figure 9). Congenital diseases in siblings with a clear phenotype present in infancy are usually indicative of inherited genetic conditions, such as in studies I and II. The sample materials from multiple affected and healthy individuals allowed the genome-wide studies on genetic causes of rare diseases in two families. In study III, pure fresh frozen tumor and myometrium tissue materials served as a basis for the molecular genetic study of uterine leiomyomas. Uterine leiomyomas display a good *in vivo* model of benign tumor growth, as multiple independent tumors can arise in single patients' uterus without malignant degeneration.

Finland and other Nordic countries have long traditions in nationwide population level registration. FCR is a unique database with complete population-based data of cancer incidences applicable for epidemiological studies (Pukkala, 2011; Teppo *et al.*, 1994). In study IV, the registry-based data were utilized to identify new tumor predisposition phenotypes (Figure 9). Because every Finnish citizen has a PIC since the 1960s, the data linkage between FCR and NPR was possible, and birth municipalities and family names at birth could be retrieved for a large number of patients in FCR. Correspondingly, the parish records are a major resource for genealogical information on births, deaths and marriages since the 16th century (Peltonen *et al.*, 1999a), which was utilized successfully in the genealogy in studies II and IV.

Traditional methods to identify Mendelian disease genes have included Sanger sequencing of positional candidate genes selected based on genetic mapping and relevance of the encoded proteins to the disease development (Figure 9). In study I, this research strategy was employed successfully, owing much to the numerous studies of left-right axis development in model organisms (Levin, 2005; Okumura *et al.*, 2008; Shen, 2007). The introduction of NGS technologies during this thesis work allowed the search for novel candidate genes in study II, and the characterization of leiomyoma genomes in study III (Figure 9).

Much work is still needed to improve sequence variant calling and delineation of disease-causing variants from NGS data. Many NGS studies have failed to detect SV calls of certain size ranges, and combination microarray- and sequencing-based approaches are needed to capture all variation in an individual genome (Pang *et al.*, 2010). Repetitive regions are problematic in alignment, and nearly 4% (~100 Mb) of the genome is considered poorly mappable with the current high-throughput and short read approaches (DePristo *et al.*, 2011; Treangen and Salzberg, 2011). A large number of false positive calls is a considerable challenge especially in SV analyses of short read sequencing. We disregarded all SV calls that originated from problematic genomic regions in study III. It is possible that many true variants have also been missed in our WGS data analyses, and for more sensitive detection of all types of variants, several tools should be applied (Pabinger *et al.*, 2014). Furthermore, assessment of disease-causing or disease-associated variants from a large number of background variants requires careful consideration in the era of NGS. In the studies of this thesis, specificity was prioritized as we tried to address the biological and etiological significance of the most prominent genetic causes of diseases in multiple individuals. False assignments are a substantial issue in disease genetics, and general recommendations have recently been given for assessment of causality of sequence variants. Most importantly, any novel gene should be implicated in disease only when genetic support from multiple independent cases can be provided (Figure 9) (MacArthur *et al.*, 2014).

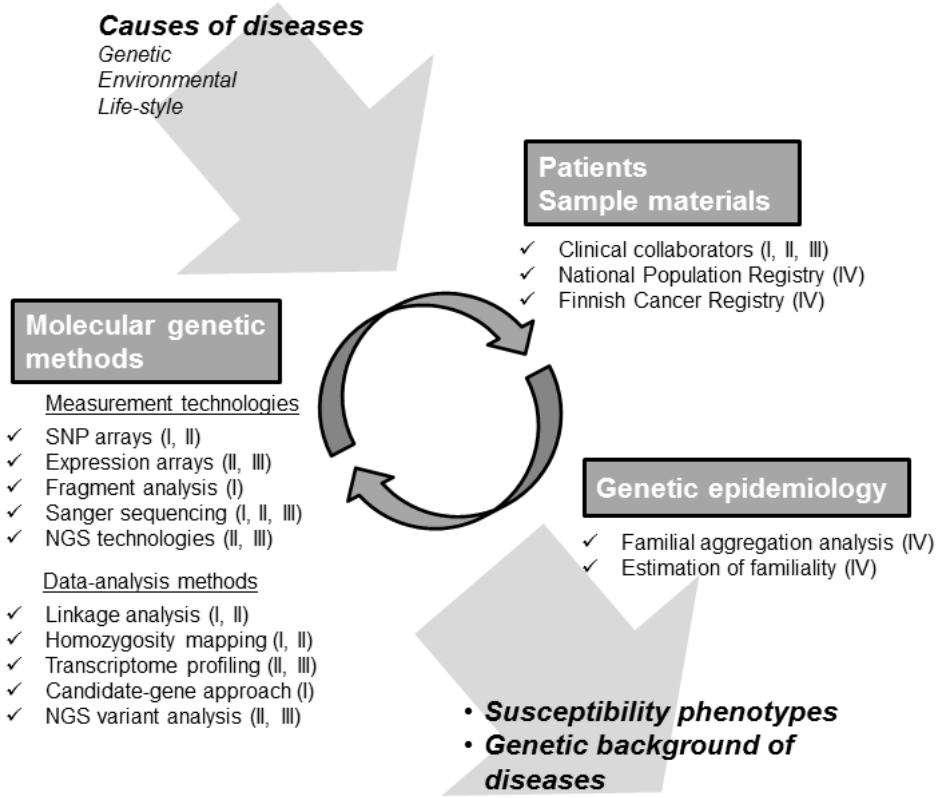

Figure 9. Summary of approaches for studying susceptibility phenotypes and genetic background of diseases. Genetic epidemiology studies the role of genetic factors in disease etiology and familial aggregation. Epidemiological measures can direct the selection of appropriate molecular genetic methods for identification of underlying genetic causes. Any genetic change should be studied in multiple idependent patient and sample materials before final assessment of causality. The materials and methods used in the studies of this thesis are indicated with roman numbers.

6.1 The role of *GDF1* in isomerism and heart defects (I)

Heterogeneity of laterality defects in human patients has complicated the identification of disease-causing genes, and no gene was unambiguously identified to follow autosomal recessive inheritance of RAI before study I was conducted. The assumption of autosomal recessive inheritance was justified in the study family, as there were five affected patients in one generation, including both males and females. Although the number of linkage-compatible regions was high after the initial genome-wide analysis, positional candidate-gene approach reduced the number of interesting loci to four. The microsatellite genotyping confirmed one genomic region as a plausible disease locus with one candidate gene, *GDF1*.

Two truncating mutations in *GDF1*, c.681C>A (p.C227X) and c.909insC, segregated with the RAI phenotype in an autosomal recessive manner in the family. The parents were healthy heterozygous carriers of one of the two mutations. Similar to the phenotype in the affected siblings in this study, *Gdf1*-/- mice have been shown to display a spectrum of laterality defects including complex cardiac defects and duplication of the right-sided pattern in the lungs (right pulmonary isomerism) (Rankin *et al.*, 2000). Partial embryonic lethality of *Gdf1*-/- mice after E14.5 has been reported in two studies (Andersson *et al.*, 2006; Rankin *et al.*, 2000). Complex heart defects reduced the viability of the *Gdf1*-/- mice, resembling the poor prognosis of human patients with RAI (Eronen *et al.*, 2013; Freedom *et al.*, 2005).

GDF1 belongs to the transforming growth factor beta (TGFβ) family of proteins with three-dimensional cystine-knot conformation, which exist mostly as homo- or heterodimers (Sun

and Davies, 1995). The mature GDF1 consists of the residues from 254 to 372 and is formed from its larger precursor preproprotein after translation by proteolytic processing (see Figure 3 in the original publication I). First the amino-terminal signal sequence directing the protein to secretion is removed, after which the propeptide is cleaved to release the active mature protein. The nonsense mutation, p.C227X, resides before the cleavage site, thus preventing the production of the mature protein, whereas the insertion, c.909insC, changes the mature protein product from the residue 303 onwards leading to a severely altered protein and truncation. After the insertion, three out of the six cysteines needed for the cystine-knot conformation are lost (see Figure 3B in the original publication I).

Nodal is known to be one of the morphogens asymmetrically localized to the left side of the developing embryo by nodal flow. This is supported by the findings that loss of Nodal in the node causes disruptions in the subsequent molecular asymmetry and left-right patterning in mice (Brennan *et al.*, 2002). *Gdf1*-/- embryos lacked expression of Nodal in the left half of early mouse embryo called left lateral plate mesoderm (LPM), suggesting that Gdf1 is required for the left-sided expression of Nodal (Rankin *et al.*, 2000; Tanaka *et al.*, 2007). Furthermore, Gdf1 has been suggested to enhance Nodal activity in the left LPM and, subsequently, to activate a known midline-specific gene, Lefty1 (Tanaka *et al.*, 2007). Altogether, previous studies seem to support the essential role of GDF1 in the early left-right axis specification during embryogenesis.

A previous report had examined 375 unrelated individuals with cardiac abnormalities for mutations in *GDF1* (Karkera *et al.*, 2007). They found eight heterozygous mutations, of which two (p.C227X and p.C267Y) were likely loss-of-function mutations. In addition to the p.C227X mutation that was also identified in our study, they found a heterozygous mutation (p.C267Y) changing the first cysteine needed in the cystine-knot formation. However, Karkera *et al*. (2007) had only studied patients but not their parents. Therefore, it is not known if any of the healthy parents were carriers of the mutation, contradicting their own suggestion that heterozygous loss-of-function mutations in *GDF1* are causative for congenital heart defects (CHD). Our results from the blood donors, CRC patients and healthy parents and their siblings, who were confirmed not to have any history of heart defects, suggest that one loss-of-function mutation in *GDF1* is not enough for the occurrence of CHD. Nevertheless, it is possible that the other six mutations found by Karkera *et al*. (2007) have dominant negative effect through dimerization or receptor binding in aberrant cardiac development. At least three of the mutations (p.S56P, p.P59T and p.A69T) reside in a proportion of GDF1 implicated in dimerization (Karkera *et al.*, 2007; Sun and Davies, 1995). Gene dosage effect as an etiologic factor has been proposed in 4.3% of CHD cases that were shown to harbor rare CNVs in genes associated with cardiac malformations. Interestingly, one large gain has been observed in *GDF1* in a CHD patient (Tomita-Mitchell *et al.*, 2012). Combination of genetic and environmental factors may be required for the manifestation of cardiac anomalies, and further work is needed to investigate the possible role of heterozygous *GDF1* mutations in CHD.

Incidences of 0.67 and 0.49 per 10,000 live births for LAI and RAI, respectively, have been reported in the Canadian population (Lim *et al.*, 2005). High rate of prenatal termination of pregnancies might lower the incidence of RAI among live births (Yan *et al.*, 2008). In our study, we identified altogether 11 carriers of heterozygous truncating *GDF1* mutations. The nonsense mutation (p.C227X) was identified in the Finnish and UK blood donors as well as in the study of Karkera *et al*. (2007), demonstrating the presence of the mutation also in the US population. Although the role of heterozygous *GDF1* mutations in CHD remains contradictory, recessively inherited mutations in *GDF1* seem clearly causative for the RAI phenotype in the study family, supported by the similar phenotype in knockout mice. Further

studies are warranted to examine the proportion of *GDF1* mutations underlying RAI, which is one of the most severe forms of congenital cardiac defects (Eronen *et al.*, 2013).

6.2 Candidate genes of novel severe intellectual disability syndrome (II)

Six patients with severe ID syndrome of unknown etiology originated from the same village from the north-east of Finland. The village is part of the late-settlement region inhabited by a well-characterized young isolate useful for identification of rare disease-causing variants (Jakkula *et al.*, 2008). Consanguinity of the healthy parents of the patients could be ascertained by in-depth genealogy, which led us to suspect autosomal recessive ID, and a single origin genetic defect. In addition to delayed psychomotor development and hypotonia, abnormalities in the visual focusing and strabismus were noted in the patients. The phenotype included also coarsening of facial features and obesity suggestive of a storage disease and a deficiency in metabolism.

Genome-wide analyses revealed a long 11.5 Mb stretch of consecutive homozygous SNP genotypes at 3p22.1-3p21.1 shared between all six patients. The locus showed a LOD score of 11, confirming the autosomal recessive linkage. The WGS data of one of the patients allowed identification of three candidate genes, *TKT*, *P4HTM* and *USP4* with potentially protein damaging sequence changes within the homozygous region. The variants were present with 0.3-0.7% allele frequencies in the population-matched controls, and only heterozygous carriers were found in the control sets examined. None of the genes were earlier shown to be mutated in ID phenotypes.

TKT, which is a ubiquitous thiamin-dependent enzyme linking the pentose phosphate pathway (PPP) and glycolysis (Schenk *et al.*, 1998) has a clear metabolism-related function. The identified missense change (p.Ile181Val in isoform 1 and p.Ile189Val in isoform 2) affects a conserved isoleucine residing in a beta sheet structure of the protein (Mitschke *et al.*, 2010). A patient with a deficient activity of RPI, another PPP enzyme, was previously described with leukoencephalopathy, psychomotor retardation, epilepsy and a slow neurological regression (MIM #608611) (Huck *et al.*, 2004). The symptoms of our patients were similar except for leukoencephalopathy. Huck *et al*. (2004) had reported compound heterozygous mutations (frameshift and missense type changes) in the *RPI* gene, decreased activity of the enzyme, and highly elevated levels of ribitol and D-arabitol in the brain and body fluids of the patient. We tested the enzyme activity of TKT from the patients' lymphoblasts, and polyol levels of the patients' urine and plasma, but could not detect consistent differences as described in the RPI deficiency. Knockout of *Tkt* has been shown to be embryonic lethal in mice (Xu *et al.*, 2002), suggesting that the patients' ID syndrome is not caused by total loss of TKT.

The variant at the splice site of *P4HTM* had the most dramatic effect on protein sequence, as the in-frame loss of exon 6 (p.Arg296Ser+Val297_Arg358del) was detected in patients' blood. P4H-TM has been reported to hydroxylate HIF-1α that is targeted for degradation in normoxic cellular conditions (Koivunen *et al.*, 2007). HIF-1α and HIF target genes are important during early development of the brain (Trollmann and Gassmann, 2009). Unlike other known prolyl 4-hydroxylases, P4H-TM has a transmembrane domain and is localized to the endoplasmic reticulum membranes. The catalytically important C-terminal region of P4H-TM contains two histidines and one aspartate that bind the Fe2+ (Koivunen *et al.*, 2007). Two of these, His328 and Asp330, are lost due to the splicing defect causing complete loss of the proposed enzyme function. Interestingly, *P4HTM* expression seems to be highest in the central nervous system (BioGPS, expression profile of the 222125_s_at probeset) (Wu *et al.*, 2009), and in the eye and brain of Zebrafish (Hyvarinen *et al.*, 2010). Although we could not

verify the suggested loss of P4H-TM activity in the patients' lymphoblasts, the tissue-specific expression of *P4HTM* could explain the patients' abnormalities in neural development and vision. The tissue-specific function of P4HTM should be studied with a mouse model, for example.

The third candidate gene, *USP4*, displayed expression of two alternative transcripts with loss of one or two amino acids (p.[Gly658del,Gly658_Glu659del]) in the patients' blood. USP4 deubiquitinates target proteins, and has been shown to regulate expression of the Adora2a receptor, a Gs-coupled receptor at the cell surface of cultured hippocampal neurons (Milojevic *et al.*, 2006). Signaling pathways that show significant enrichment of differentially expressed genes may provide evidence of the underlying genetic defect. Our expression analyses showed altered expression of genes that were enriched in the CREB, G-protein coupled receptor and cAMP signaling pathways. The three pathways regulate CREB transcription factors through cAMP signal transduction, and the enrichment could be related to the altered function of USP4 in regulating Gs-coupled receptors. CREB is a known player in certain neurodegenerative diseases, such as in Coffin-Lowry syndrome caused by mutations in the X-chromosomal CREB protein kinase *RPS6KA3* (MIM #303600) (Trivier *et al.*, 1996) and in Rubinstein-Taybi syndrome caused mainly by *de novo* heterozygous mutations in *CREB binding protein* (MIM #180849) (Petrij *et al.*, 1995). We could not detect alterations in the cAMP concentrations or in the CREB protein levels in the patients' lymphoblasts, although the effect might be specific to CREB5, which was slightly downregulated in the expression analysis (p-value=0.08). Unfortunately, we could not get CREB5 antibodies to work in immunoblotting.

Altogether, the novel ID syndrome might be a monogenic, digenic or trigenic disease specific to the haplotype of the three variants in homozygous form, although the possibility of a causative noncoding change in the 3p22.1-3p21.1 locus cannot be excluded based on this study. The frequency of the identified candidate variants in the north-east of Finland suggests that more patients with the same haplotype are likely present in the region. Additional fifteen patients with varying degree of compatibility with the ID patients of the study family were examined for the three variants with negative results. Further studies with a larger set of patients are warranted to clarify the role of the identified candidate genes in the development of severe ID.

6.3 Genetic changes in development of uterine leiomyomas (III)

The application of NGS and microarray technologies allowed us to characterize genomic alterations across a wide range of uterine leiomyomas. We selected tumors without *MED12* and *FH* mutations, as well as *MED12*-mutant and *FH*-deficient tumors, for the experiment. Somatic point mutations affecting protein coding genes were rare in the uterine leiomyomas. However, CCR events were detected in 15 tumors. Some CCR events had created 20 or more breakpoints in a single chromosome, while other CCR events displayed three breaks that all had occurred simultaneously. CCRs were inferred computationally from chained event graphs utilizing SV and CNA calls. Step-wise accumulation of DNA breakage and anomalous repair of break-end pairs seems an unlikely mechanism, as multiple repaired breakpoint junctions would have to be rebroken to form such interconnected rearrangements. The CCR events in the leiomyomas resembled the phenomenon of chromothripsis that was earlier associated with advanced cancers (Rausch *et al.*, 2012; Stephens *et al.*, 2011).

Chromothripsis was first described as a catastrophic shattering of one or two chromosomes, and as a formation of clustered rearrangements and focal losses of DNA (Stephens *et al.*, 2011). The most likely mechanisms to explain chromothripsis include replication stress

caused by activated oncogenes at the G1/S phase of the cell cycle and mitotic errors (Jones and Jallepalli, 2012). Whether or not the CCR events in leiomyomas are formed through the same mechanism as chromothripsis remains to be studied. Compatible with the chromothripsis phenomenon, one leiomyoma with the most breakpoints displayed a CCR event causing activation of the *CCND1* gene, which is a known regulator of the cell cycle G1/S transition. Although chromothripsis has been associated with *TP53* mutations (Northcott *et al.*, 2012; Rausch *et al.*, 2012; Wu *et al.*, 2012), we detected only one tumor with a *TP53* rearrangement that was created through a CCR event and no point mutations in *TP53* in any of the leiomyomas studied. In some tumors, we could show that separate CCR events had occurred multiple times in a single tumor cell lineage. After the publication of study III, prostate tumors were also shown to harbor sequential CCR events similar to ours. The CCR events in a subset of prostate tumors also affected more than one or two chromosomes, and involved fewer breakpoints than detected in chromothripsis (Baca *et al.*, 2013).

CCRs seem to be a major cause of chromosomal alterations creating tissue-specific changes in tumors lacking *MED12* and *FH* mutations (Figure 10). CCRs had produced the known rearrangements between the *HMGA2* and *RAD51B* loci, or alterations of the *COL4A5*/*COL4A*6 locus in multiple tumors. CCRs are not expected to create these types of selective changes with such a high frequency as detected in the tumors in our WGS experiment; thus, CCR events must not be rare in precursor myometrial cells. Majority of the most dramatic remodeling events likely leads to apoptosis (Figure 10).

Gene expression data could confirm oncogenic *HMGA2* and *HMGA1*activation in all eight tumors with the upstream breakpoints of the respective genes. Similar to a previous study (Hodge *et al.*, 2012), genes involved in the G1/S transition were dysregulated among *HMGA2*/*HMGA1* overexpressing samples. The most frequent translocation partner, *RAD51B* was shown to provide *HMGA2* with an effective enhancer. RAD51B is a DNA repair protein that is disrupted in leiomyomas by the rearrangements, and may have a role in maintaining the genome stability in myometrial cells.

Three samples with the alterations of *COL4A5*/*COL4A*6 displayed a clear upregulation of *IRS4*. *COL4A5* is located head-to-head with *COL4A*6 at Xq22.3, and the two genes share a bidirectional promoter region (Khoshnoodi *et al.*, 2008). Mutations in the *COL4A5* gene cause Alport syndrome (MIM #301050), whereas only partial deletions of the promoter region have been reported in patients with Alport syndrome and diffuse leiomyomatosis characterized by smooth muscle overgrowth (ATS-DL, MIM #308940). Activation of a neighboring gene of the *COL4A5*/*COL4A*6 locus has been proposed as a mechanism for the smooth muscle overgrowth in ATS-DL (Thielen *et al.*, 2003). Sporadic deletions at the *COL4A5*/*COL4A*6 locus have earlier been detected in leiomyomas of esophagus (Heidet *et al.*, 1998), and now in our study. The *IRS4* gene, encoding insulin receptor substrate 4, is a plausible candidate for smooth muscle overgrowth, since dysregulation of insulin-like growth factor signaling has been earlier implicated in leiomyomas (Burroughs *et al.*, 2002).

Chromosome arm 7q has been shown to display frequent losses in leiomyomas, with a few candidate target genes identified earlier (Ligon *et al.*, 2002; Quintana *et al.*, 1998). We could show specific rearrangement breakpoints in the *CUX1* gene located in the minimally deleted region across our tumors. In one tumor, biallelic inactivation by two independent CCR events was observed, suggesting that *CUX1* is a tumor suppressor. CUX1 functions as a transcription factor, regulating expression of DNA damage response genes (Vadnais *et al.*, 2012). At the same time of our study, *CUX1* was independently reported as a specific target gene for haploinsufficient inactivation in acute myeloid leukemias and in uterine leiomyomas

(McNerney *et al.*, 2013; Schoenmakers *et al.*, 2013). Disruptions of *CUX1* co-occur with *MED12* mutations and *HMGA2* activation as shown by us and others (Markowski *et al.*, 2012), indicating that *CUX1* inactivation is a secondary change in the development of uterine leiomyomas.

Tumors clustered separately according to the alterations of *MED12*, *FH*, *HMGA2*/*HMGA1* or *COL4A5*/*COL4A*6 in the gene expression analysis, with the exception of a single tumor displaying the massive chromothripsis, *CCND1* upregulation and *COL4A5*/*COL4A*6 alteration. The clustering has also been confirmed with our recent data including more tumors (Mehine *et al*. unpublished results). Thus, it seems that the development of uterine leiomyomas is driven by at least four mututally exclusive molecular pathways having distinguishable global expression profiles. Majority of leiomyomas are formed through *MED12* mutations (Figure 10), which seem to lead to development of smaller tumors (Mäkinen *et al.*, 2011b). Tumors displaying *HMGA2* activation represent the majority of cytogenetically abnormal and large leiomyomas (Bulun, 2013) formed through CCRs or simple balanced translocations (Figure 10).

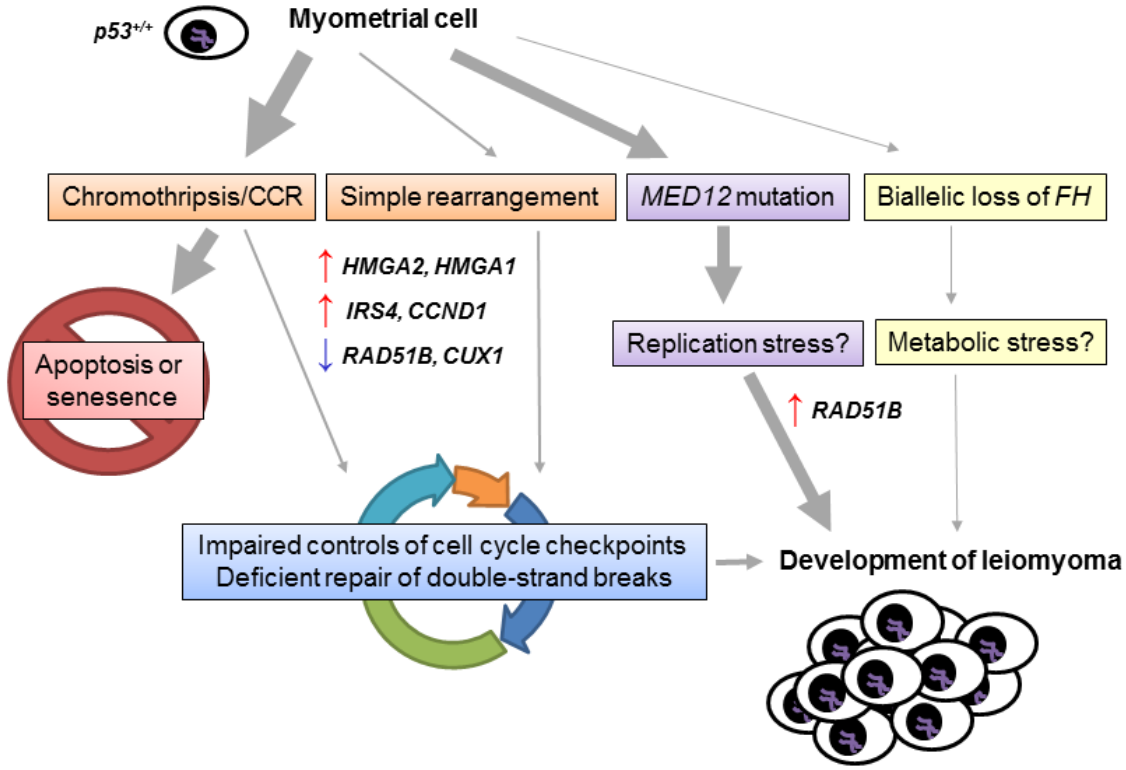

Figure 10. Model of uterine leiomyoma development. Different mutational mechanisms create changes that affect the normal function of myometrial cells. Somatic *MED12* mutations drive the majority of lesions. The upregulation of *RAD51B* in *MED12* mutated tumors might reflect cells responses to replication stress. A substantial proportion arises through complex chromosomal rearrangements (CCRs) that impair the control of cell cycle checkpoints and repair of DNA double-strand breaks, such as rearrangements between the *HMGA2* and *RAD51B* loci. Simple chromosomal rearrangements also create similar changes. A less frequent route is inactivation of *FH*, resulting in the tumor development possibly through metabolic stress.

Sequence analysis was also able to provide definitive evidence for the clonal origin of multiple tumors in three patients. Similar phenomenon had also been proposed earlier by cytogenetic analyses (Nibert and Heim, 1990; Nilbert *et al.*, 1990). We were able to show that subclonal cells from one tumor had disseminated a secondary tumor within a single uterus.

Further studies are needed to clarify, what prevents tumor cells from metastasizing, and whether cell lineages with a specific genetic background are more prone to seed multiple tumors.

6.4 Identification of tumor susceptibility phenotypes using registry-based data (IV)

Familial aggregation studies of cancer are usually restricted to patients whose family members are known. We inferred family relations from the birth municipality and family name at birth combinations for all the patients in FCR; thus we were able cover remarkably more cases and detailed tumor types than previous comprehensive studies conducted in the Icelandic or Utah populations (Albright F Ph *et al.*, 2012; Amundadottir *et al.*, 2004; Goldgar *et al.*, 1994). We were able to estimate the familiality of both common and rare tumor types by developing the cluster score method. The cluster score ranking highlighted known rare cancer predisposition syndromes such as Von Hippel-Lindau syndrome (VHL, MIM #193300) and Multiple Endocrine Neoplasia type 2 (MEN2A, MIM #171400, and MEN2B, MIM #162300), providing evidence that our method is appropriate for familiality estimation.

Also tumor types without known predisposition phenotype were featured in the cluster score ranking, including Kaposi sarcoma. Thorough genealogy work of the MN-clusters of KS patients identified seven families with two first-degree relatives with KS, and a remarkable family with five KS cases. These confirmed familial cases of KS represented 70% of the cases in the MN-clusters, most of which were high confidence clusters in terms of O/E and CI-low values. Thus, it seems that our proportional incidence and statistical significance measures are appropriate for identifying clusters with true relatives. The systematic registry-based clustering has been performed yearly after the publication of study IV, and one additional family with two first-degree relatives with KS has been identified.

The success of the clustering method owes much to the diversity of family names in Finland and to the isolation of villages in the history of the Finnish population. Patients with common family names in Finland were less likely to produce statistically significant clusters. Tumor types that occur with higher frequency in sibships and in father-child pairs are overrepresented in our clusters, because family name at birth is usually inherited from the paternal side in Finland. Unfortunately, systematic search for true family relationships from NPR is limited to only those patients who have lived in the 1960s or later. Both or one parent was found from NPR for only ~15% of patients in the clusters, when this systematic approach was tested with selected tumor types.

Familial relations of the KS cases in N-clusters were not studied thoroughly due to large numbers of cases and the slowness of genealogy work. A number of false positives are expected among these clusters, although N-clusters are not constrained by municipality borders. Therefore, we have now considered dividing the Finnish municipalities into regions and performing clustering based on tumor type, birth place regions and family names at birth.

In addition to KS, small intestine neuroendocrine tumors showed strong familial aggregation in our study. High familial risk has been reported in other studies as well (Hemminki and Li, 2001; Hiripi *et al.*, 2009; Kharazmi *et al.*, 2013), but genetic basis of the predisposition is largely unknown. Papillary thyroid adenocarcinoma, chronic lymphatic leukemia, and squamous cell carcinoma of lip, all considered here as common tumor types with over 7,000 registered patients in FCR, showed also high ranking according to cluster score. Interestingly, the same cancer sites have high incidence in familial cases who are distantly related to each other (more than first cousins), as demonstrated in a study of the Utah population (Albright F

Ph *et al.*, 2012). Our method should also be able to detect relatives outside the nuclear family, but genealogy work is needed to confirm family relationships.

We have identified so far 21 familial KS cases, and 74 familial cases have been reported elsewhere, mostly in the Jewish population (Almohideb *et al.*, 2013). The Finnish family of five KS cases is among the largest families reported to date (Cottoni *et al.*, 1996; DiGiovanna and Safai, 1981). HHV8 infection is required for the manifestation of KS, and persons with immunodeficiency are at higher risk for KS (Safai *et al.*, 1985). KS incidence and HHV8 prevalence varies in different populations (Mesri *et al.*, 2010), Finnish showing an average standardized KS incidence rate of 0.1-0.2 per 100,000 person-years. As shown in our study, the incidence of KS according to municipality of birth varied strongly between Eastern and Western Finland, although HHV8 infection seems to be rare in both regions. HHV8 prevalence should be studied more carefully in order to know whether the uneven distribution of cases is due to uneven distribution of HHV8 prevalence or genetic susceptibility in Finland. Intrafamilial transmission of HHV8 has been suggested (Mancuso *et al.*, 2011), which may explain some of the familial cases observed in our study. The five KS cases in a single family had seemingly normal immunity, indicating that HHV8 infection must be accompanied with genetic susceptibility to develop tumors in the family.

7. Conclusions and future prospects

Epidemiological and molecular approaches in studies I-IV led to characterization of novel susceptibility phenotypes, and genetic background of right atrial isomerism, severe intellectual disability, and uterine leiomyomas.

Study I: RAI was recessively inherited in a family with five affected siblings and healthy parents. Positional candidate-gene approach allowed identification of two truncating mutations in *GDF1*, shown to be compound heterozygous in the patients. Similar phenotype had been characterized in *Gdf1* knockout mice, providing firm evidence that the mutations in *GDF1* are causative for RAI in the family. We were able to identify multiple healthy heterozygous carriers of truncating *GDF1* mutations, which bring into question the role of heterozygous loss-of-function mutations in *GDF1* in congenital cardiac defects as previously reported. The proportion and type of *GDF1* mutations underlying laterality and congenital cardiac defects should be studied in more patients. Screening of *GDF1* mutations is feasible for molecular diagnosis of RAI at earlier stages of pregnancy than echocardiography, and could facilitate genetic counseling for families with a history of laterality defects.

Study II: A novel autosomal recessive ID syndrome was characterized and an unambiguous linkage was found at the 3p22.1-3p21.1 locus displaying a large homozygous region in six consanguineous ID patients from four sibships. Whole-genome sequencing pinpointed three genes with potentially protein damaging sequence changes as novel candidates for syndromic ID. No definite conclusion of the role of the genes could be drawn. Clinical and genetic data obtained in this study can be used to identify patients with a similar phenotype elsewhere. Further work is obviously needed to show the combined or individual role of the candidate genes in the etiology of the autosomal recessive ID syndrome.

Study III: Complex chromosomal rearrangements resembling chromothripsis were detected in uterine leiomyomas by whole-genome sequencing. Although earlier associated with malignant tumors, the massive remodeling phenomenon can also occur in these benign tumors and create driver changes that promote leiomyoma growth in sequential fashion. CCRs were common in tumors without *MED12* and *FH* mutations, warranting further studies on mechanisms underlying CCRs and chromothripsis. Four molecular subgroups driven by alterations of *MED12*, *FH*, *HMGA2*/*HMGA1* or *COL4A5*/*COL4A*6 were characterized, and in some patients multiple separate tumors displayed common clonal origin. Leiomyomas with distinct biological properties should be studied further to allow personalized treatment of patients with more severe symptoms and aggressive tumor growth.

Study IV: Measures of familiality were successfully derived for both rare and common tumor types by utilizing the population based data in the Finnish Cancer Registry. The systematic registry-based search for familial aggregation of all types of cancers was able to identify familial cases suitable for sample collection, and for further research on underlying genetic predisposition. Kaposi sarcoma showed strong familiality, and prevalence of viral and genetic predisposition in Finland should be studied further, as an unequal distribution of Kaposi sarcoma patients in Eastern and Western Finland was noted. Furthermore, FCR can be used to derive sample information for all the patients examined in pathology laboratories, enabling genetic studies of the novel susceptibility phenotypes with the advanced genome-wide approaches.

The approaches used in this thesis are applicable to studies of a variety of human diseases. The new measurement technologies permit the unbiased genome-wide analyses of variation, although large numbers of samples from patients with similar phenotypes are required to

determine unequivocally the role of genetic changes in diseases. Availability of patient and kinship information is increasing as electronic population-based registries such as NPR and FCR were established around 50-60 years ago in Finland, which should be harnessed for future studies of genetic epidemiology. Often knowledge of the underlying genetic changes is not sufficient, and experimental approaches are needed. There it becomes more and more important to understand specific biological systems in relation to disease manifestation. Joint effect of environment and genes should also be taken into account in future etiological studies of complex diseases.

8. Acknowledgements

This work was carried out in the Tumor Genomics Research Group at the Department of Medical Genetics and Research Programs Unit, University of Helsinki, during 2009-2014. The present and former heads of the Department of Medical Genetics and Research Programs Unit are thanked for providing research facilities and environment for high quality research.

I am grateful to my supervisor Academy Professor Lauri Aaltonen for the many years I have been privileged to work in his cutting-edge research group and to be part of so many interesting research projects. I could not have hoped for a better environment for me to grow as a researcher and to learn how to do competitive research. There is no rival to Lauri's commitment to supervising (also in matters outside the research life) and taking care of everyone's well-being, which I appreciate a lot. Thank you for all the valuable discussions we've had, and for wasting your sense of humor on me. My other supervisor, Esa Pitkänen, is thanked for his continuous support in understanding various computational approaches and in developing new ideas. Esa seems to have a never-ending forbearance for listening and problem solving, which makes him always easy to approach, even in the worst moments of doubt. Your presence has lightened my mood so many times.

I am sincerely thankful for my thesis reviewers and thesis committee members, Professor Matti Nykter, and Docents Marjo Kestilä and Janne Nikkilä, for their time and knowledge that they have given to my thesis work. Your advice and comments have been indispensable for the maturation of this book.

I want to express my warm gratitude to all the co-authors and collaborators of the studies in this thesis. Professors Kristiina Aittomäki and Jukka-Pekka Mecklin, Docents Marianne Eronen, Eero Kajantie, Jukka Moilanen, Ralf Bützow and Jari Sjöberg, doctors Elisa Rahikkala and Leila Pajunen, and Sirpa Miettinen have formed the basis for the studies in the clinic by characterizing the patients, collecting sample materials and sharing their invaluable medical expertise. Professor Eero Pukkala, Miia Artama, Toni Patama and Hilkka Laasanen have enabled the sensible use and interpretation of the Finnish Cancer Registry data. Professors Peppi Karppinen and Johanna Myllyharju, Mirjam Wamelink, Massimiliano Gentile, Paul Knekt and Maarit Laaksonen have provided their in-depth knowledge related to the individual study questions.

My sincere gratitude goes to the present and former post-docs in the Aaltonen lab: Auli, Rainer, Esa, Pia, Sari, Heli, Outi, Javier, Kimmo, Niko and Tuomas. Auli and Rainer, you have taught and closely guided me from the beginning of my thesis work and throughout the years, for which I am grateful. Pia and Outi, thank you for your encouragement in my thesis work, and vision in the leiomyoma and tumor gene projects. Sari, you were such a great instructor for me in my first summer in the lab, and I am particularly thankful that you have shared your honest anxiety with me in matters such as karonkka. Javier, thank you for taking care of the maintenance of IT related stuff for so many years. Kimmo and Niko, I appreciate your computational skills and I am grateful that I have got to work with you in some of the most recent projects. The Lehtonen couple, Heli and Rainer, special thanks for your reliable, bright and cheerful company.

The incubators, Mervi, Miika, Riku, Heikki R., Heikki M., Yilong, Iikki, Johanna, Manuel and Elina, have shared the inspiring office space and thoughts with me along the way. I have learned to know each and every one of you personally, for which I am deeply grateful. Miika, I appreciate your enthusiasm and ideas for the leiomyoma research and for saving "millions of women worldwide". Riku, you have done so much unselfish work for the next-generation

sequencing projects and for the Aaltonen publishing style which cannot be acknowledged enough. Heikki R. and M., working with you in any pipeline or IT related task is always fun and reassuring. Yilong, I am happy that I had you as my close friend since we were still studying genetics and bioinformatics; I have learned so much from you. Iikki, the newest member of Incubator, thank you for taking me out for beers when I had the most pressing time with completing the thesis work. Your outspokenness is always refreshing. Johanna, my fellow summer student in the lab in 2007, I am grateful to your loyal friendship during all these years we have been in the Aaltonen group. I am particularly thankful for your companion in the Nordic cancer epidemiology course and Berlin trip, being my personal doctor (even though you wouldn't like to be), being my training partner with your extreme endurance, making me laugh with your great stories and sense of humor, and most importantly, for always passing me your drinks.

My partner in crime, Mervi, has been a priceless component of this thesis work; she is the one I first share my thoughts with, and who I count on reading and commenting my texts, among other things. Mervi has a wide knowledge-base and an efficient "k-mer data retrieval" in her mind for any scientific or otherwise compelling discussion. Mervi's cheerful personality and ready-to-go-anywhere attitude have kept me afloat during these years. Thanks for bearing with me even on holidays. I especially want to thank you for the countless times you have encouraged me and translated my thoughts.

My warm appreciation goes to the other present and former fellow students in the Aaltonen lab: Netta, Kati, Hanna, Silva, Alex, Tatiana, Iina T., Anniina, Ulrika, Tomas, Hande, Jaana and Taru. Netta, Kati and Hanna are specially thanked for their skillful cooperation in the leiomyoma research. Iina T. and Anniina showed me great example in the early years of my thesis work. Alex, thank you so much for taking us to Arosa, and Tatiana, thanks a lot for showing us Menorca. Both trips were such unique, unforgettable experiences in the best company. Netta and Silva, I value your determination and distinctive characters, and I am glad that I have got to work closely with both of you and to know you better in our get-togethers.

I am deeply thankful for Sini, Sirpa and Marjo for the excellent assistance, genealogy work, and taking care of the sample deliveries, and for Inga-Lill, Iina V., Mairi and Alison for their laboratory expertise. My thesis work has depended on your contributions on numerous ways. Sini and Inga-Lill are specially thanked for keeping the lab in order with their long-term experience. Sini, above all, is the essence of the social activities and cheerful atmosphere which make the Aaltonen group such a great place to work in.

The other colleagues in the neighboring labs, Center of Excellence in Cancer Genetics and Genome-Scale Biology program are thanked for interesting scientific discussions and collaborative projects. I also want to thank Denis, Maral and Mervi for the great and dynamic company in organizing the CancerBio Summer school and Amazing Cancer Race.

I am indebted to the families and individual patients for participating in the studies.

My sincere appreciation goes to Tiia Pelkonen for reviewing the language of my thesis.

Finally, I want to thank my parents, Raija and Jouko, my siblings and their families, and my friends outside the research life, especially my close friend Tilli, for continuous understanding and support.

Helsinki Graduate Program in Biotechnology and Molecular Biology (GPBM) and Integrative Life Science (ILS) Doctoral Program are thanked for funding my thesis work, two conference trips and a student excursion to Japan. The courses and events organized by the programs have helped me to get more perspective on research and social networks.

Helsinki, September 2014

Eevi Kaasinen

9. References

The Human Genome Project. 2014. http://www.genome.gov/10001772; last updated March 18, 2014; visited May 2014

1000 Genomes Project Consortium, Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E. & McVean, G. A. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467:** 1061-1073.

1000 Genomes Project Consortium, Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T. & McVean, G. A. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491:** 56-65.

Aaltonen, L., Johns, L., Järvinen, H., Mecklin, J. P. & Houlston, R. 2007. Explaining the familial colorectal cancer risk associated with mismatch repair (MMR)-deficient and MMR-stable tumors. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research* **13:** 356-361.

Abecasis, G. R., Cherny, S. S., Cookson, W. O. & Cardon, L. R. 2002. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics* **30:** 97-101.

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S. & Sunyaev, S. R. 2010. A method and server for predicting damaging missense mutations. *Nature Methods* **7:** 248-249.

Alam, N. A., Bevan, S., Churchman, M., Barclay, E., Barker, K., Jaeger, E. E., Nelson, H. M., Healy, E., Pembroke, A. C., Friedmann, P. S., Dalziel, K., Calonje, E., Anderson, J., August, P. J., Davies, M. G., Felix, R., Munro, C. S., Murdoch, M., Rendall, J., Kennedy, S., Leigh, I. M., Kelsell, D. P., Tomlinson, I. P. & Houlston, R. S. 2001. Localization of a gene (MCUL1) for multiple cutaneous leiomyomata and uterine fibroids to chromosome 1q42.3-q43. *American Journal of Human Genetics* **68:** 1264-1269.

Albright, F., Teerlink, C., Werner, T.L. & Cannon-Albright, L.A. 2012. Significant evidence for a heritable contribution to cancer predisposition: a review of cancer familiality by site. *BMC Cancer* **12:** 138.

Almohideb, M., Watters, A. K. & Gerstein, W. 2013. Familial classic Kaposi sarcoma in two siblings: case report and literature review. *Journal of Cutaneous Medicine and Surgery* **17:** 356-361.

Amundadottir, L. T., Thorvaldsson, S., Gudbjartsson, D. F., Sulem, P., Kristjansson, K., Arnason, S., Gulcher, J. R., Bjornsson, J., Kong, A., Thorsteinsdottir, U. & Stefansson, K. 2004. Cancer as a complex phenotype: pattern of cancer distribution within and beyond the nuclear family. *PLoS Medicine* **1:** e65.

Andersson, O., Reissmann, E., Jornvall, H. & Ibanez, C. F. 2006. Synergistic interaction between Gdf1 and Nodal during anterior axis development. *Developmental Biology* **293:** 370-381.

Baca, S. C., Prandi, D., Lawrence, M. S., Mosquera, J. M., Romanel, A., Drier, Y., Park, K., Kitabayashi, N., MacDonald, T. Y., Ghandi, M., Van Allen, E., Kryukov, G. V., Sboner, A., Theurillat, J. P., Soong, T. D., Nickerson, E., Auclair, D., Tewari, A., Beltran, H., Onofrio, R. C., Boysen, G., Guiducci, C., Barbieri, C. E., Cibulskis, K., Sivachenko, A., Carter, S. L., Saksena, G., Voet, D., Ramos, A. H., Winckler, W., Cipicchio, M., Ardlie, K., Kantoff, P. W., Berger, M. F., Gabriel, S. B., Golub, T. R., Meyerson, M., Lander, E. S., Elemento, O., Getz, G., Demichelis, F., Rubin, M. A. & Garraway, L. A. 2013. Punctuated evolution of prostate cancer genomes. *Cell* **153:** 666-677.

Badano, J. L. & Katsanis, N. 2002. Beyond Mendel: an evolving view of human genetic disease transmission. *Nature Reviews.Genetics* **3:** 779-789.

Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J. & Eichler, E. E. 2001. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Research* **11:** 1005-1017.

Bamford, R. N., Roessler, E., Burdine, R. D., Saplakoglu, U., dela Cruz, J., Splitt, M., Goodship, J. A., Towbin, J., Bowers, P., Ferrero, G. B., Marino, B., Schier, A. F., Shen, M. M., Muenke, M. & Casey, B. 2000. Loss-of-function mutations in the EGF-CFC gene CFC1 are associated with human left-right laterality defects. *Nature Genetics* **26:** 365-369.

Bell, C. J., Dinwiddie, D. L., Miller, N. A., Hateley, S. L., Ganusova, E. E., Mudge, J., Langley, R. J., Zhang, L., Lee, C. C., Schilkey, F. D., Sheth, V., Woodward, J. E., Peckham, H. E., Schroth, G. P., Kim, R. W. & Kingsmore, S. F. 2011. Carrier testing for severe childhood recessive diseases by next-generation sequencing. *Science Translational Medicine* **3:** 65ra4.

Benjamini, Y. & Hochberg, Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society.Series B (Methodological)* **57:** pp. 289-300.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., Irving, L. J., Karbelashvili, M. S., Kirk, S. M., Li, H., Liu, X., Maisinger, K. S., Murray, L. J., Obradovic, B., Ost, T., Parkinson, M. L., Pratt, M. R., Rasolonjatovo, I. M., Reed, M. T., Rigatti, R., Rodighiero, C., Ross, M. T., Sabot, A., Sankar, S. V., Scally, A., Schroth, G. P., Smith, M. E., Smith, V. P., Spiridou, A., Torrance, P. E., Tzonev, S. S., Vermaas, E. H., Walter, K., Wu, X., Zhang, L., Alam, M. D., Anastasi, C., Aniebo, I. C., Bailey, D. M., Bancarz, I. R., Banerjee, S., Barbour, S. G., Baybayan, P. A., Benoit, V. A., Benson, K. F., Bevis, C., Black, P. J., Boodhun, A., Brennan, J. S., Bridgham, J. A., Brown, R. C., Brown, A. A., Buermann, D. H., Bundu, A. A., Burrows, J. C., Carter, N. P., Castillo, N., Chiara E Catenazzi, M., Chang, S., Neil Cooley, R., Crake, N. R., Dada, O. O., Diakoumakos, K. D., Dominguez-Fernandez, B., Earnshaw, D. J., Egbujor, U. C., Elmore, D. W., Etchin, S. S., Ewan, M. R., Fedurco, M., Fraser, L. J., Fuentes Fajardo, K. V., Scott Furey, W., George, D., Gietzen, K. J., Goddard, C. P., Golda, G. S., Granieri, P. A., Green, D. E., Gustafson, D. L., Hansen, N. F., Harnish, K., Haudenschild, C. D., Heyer, N. I., Hims, M. M., Ho, J. T., Horgan, A. M., Hoschler, K., Hurwitz, S., Ivanov, D. V., Johnson, M. Q., James, T., Huw Jones, T. A., Kang, G. D., Kerelska, T. H., Kersey, A. D., Khrebtukova, I., Kindwall, A. P., Kingsbury, Z., Kokko-Gonzales, P. I., Kumar, A., Laurent, M. A., Lawley, C. T., Lee, S. E., Lee, X., Liao, A. K., Loch, J. A., Lok, M., Luo, S., Mammen, R. M., Martin, J. W., McCauley, P. G., McNitt, P., Mehta, P., Moon, K. W., Mullens, J. W., Newington, T., Ning, Z., Ling Ng, B., Novo, S. M., O'Neill, M. J., Osborne, M. A., Osnowski, A., Ostadan, O., Paraschos, L. L., Pickering, L., Pike, A. C., Pike, A. C., Chris Pinkard, D., Pliskin, D. P., Podhasky, J., Quijano, V. J., Raczy, C., Rae, V. H., Rawlings, S. R., Chiva Rodriguez, A., Roe, P. M., Rogers, J., Rogert Bacigalupo, M. C., Romanov, N., Romieu, A., Roth, R. K., Rourke, N. J., Ruediger, S. T., Rusman, E., Sanches-Kuiper, R. M., Schenker, M. R., Seoane, J. M., Shaw, R. J., Shiver, M. K., Short, S. W., Sizto, N. L., Sluis, J. P., Smith, M. A., Ernest Sohna Sohna, J., Spence, E. J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C. L., Turcatti, G., Vandevondele, S., Verhovsky, Y., Virk, S. M., Wakelin, S., Walcott, G. C., Wang, J., Worsley, G. J., Yan, J., Yau, L., Zuerlein, M., Rogers, J., Mullikin, J. C., Hurles, M. E., McCooke, N. J., West, J. S., Oaks, F. L., Lundberg, P. L., Klenerman, D., Durbin, R. & Smith, A. J. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456:** 53-59.

Berget, S. M. 1995. Exon recognition in vertebrate splicing. *The Journal of Biological Chemistry* **270:** 2411-2414.

Botstein, D., White, R. L., Skolnick, M. & Davis, R. W. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics* **32:** 314-331.

Brennan, J., Norris, D. P. & Robertson, E. J. 2002. Nodal activity in the node governs left-right asymmetry. *Genes & Development* **16:** 2339-2344.

Bulun, S. E. 2013. Uterine fibroids. *The New England Journal of Medicine* **369:** 1344-1355.

Burroughs, K. D., Howe, S. R., Okubo, Y., Fuchs-Young, R., LeRoith, D. & Walker, C. L. 2002. Dysregulation of IGF-I signaling in uterine leiomyoma. *The Journal of Endocrinology* **172:** 83-93.

Byun, M., Abhyankar, A., Lelarge, V., Plancoulaine, S., Palanduz, A., Telhan, L., Boisson, B., Picard, C., Dewell, S., Zhao, C., Jouanguy, E., Feske, S., Abel, L. & Casanova, J. L. 2010. Whole-exome sequencing-based discovery of STIM1 deficiency in a child with fatal classic Kaposi sarcoma. *The Journal of Experimental Medicine* **207:** 2307-2312.

Byun, M., Ma, C. S., Akcay, A., Pedergnana, V., Palendira, U., Myoung, J., Avery, D. T., Liu, Y., Abhyankar, A., Lorenzo, L., Schmidt, M., Lim, H. K., Cassar, O., Migaud, M., Rozenberg, F., Canpolat, N., Aydogan, G., Fleckenstein, B., Bustamante, J., Picard, C., Gessain, A., Jouanguy, E., Cesarman, E., Olivier, M., Gros, P., Abel, L., Croft, M., Tangye, S. G. & Casanova, J. L. 2013. Inherited human OX40 deficiency underlying classic Kaposi sarcoma of childhood. *The Journal of Experimental Medicine* **210:** 1743-1759.

Camcioglu, Y., Picard, C., Lacoste, V., Dupuis, S., Akcakaya, N., Cokura, H., Kaner, G., Demirkesen, C., Plancoulaine, S., Emile, J. F., Gessain, A. & Casanova, J. L. 2004. HHV-8-associated Kaposi sarcoma in a child with IFNgammaR1 deficiency. *The Journal of Pediatrics* **144:** 519-523.

Casey, B. 1998. Two rights make a wrong: human left-right malformations. *Human Molecular Genetics* **7:** 1565-1571.

Catherino, W. H., Parrott, E. & Segars, J. 2011. Proceedings from theNational Institute of Child Health and Human Development conference on the Uterine Fibroid Research Update Workshop. *Fertility and Sterility* **95:** 9-12.

Chang, Y., Cesarman, E., Pessin, M. S., Lee, F., Culpepper, J., Knowles, D. M. & Moore, P. S. 1994. Identification of herpesvirus-like DNA sequences in AIDS-associated Kaposi's sarcoma. *Science* **266:** 1865-1869.

Chen, K., Wallis, J. W., McLellan, M. D., Larson, D. E., Kalicki, J. M., Pohl, C. S., McGrath, S. D., Wendl, M. C., Zhang, Q., Locke, D. P., Shi, X., Fulton, R. S., Ley, T. J., Wilson, R. K., Ding, L. & Mardis, E. R. 2009. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature Methods* **6:** 677-681.

Cheng, S. K., Olale, F., Bennett, J. T., Brivanlou, A. H. & Schier, A. F. 2003. EGF-CFC proteins are essential coreceptors for the TGF-beta signals Vg1 and GDF1. *Genes & Development* **17:** 31-36.

Church, D. M., Schneider, V. A., Graves, T., Auger, K., Cunningham, F., Bouk, N., Chen, H. C., Agarwala, R., McLaren, W. M., Ritchie, G. R., Albracht, D., Kremitzki, M., Rock, S., Kotkiewicz, H., Kremitzki, C., Wollam, A., Trani, L., Fulton, L., Fulton, R., Matthews, L., Whitehead, S., Chow, W., Torrance, J., Dunn, M., Harden, G., Threadgold, G., Wood, J., Collins, J., Heath, P., Griffiths, G., Pelan, S., Grafham, D., Eichler, E. E., Weinstock, G., Mardis, E. R., Wilson, R. K., Howe, K., Flicek, P. & Hubbard, T. 2011. Modernizing reference genome assemblies. *PLoS Biology* **9:** e1001091.

Cohen, M. S., Anderson, R. H., Cohen, M. I., Atz, A. M., Fogel, M., Gruber, P. J., Lopez, L., Rome, J. J. & Weinberg, P. M. 2007. Controversies, genetics, diagnostic assessment, and outcomes relating to the heterotaxy syndrome. *Cardiology in the young* **17 Suppl 2:** 29-43.

Cooper, D. N., Krawczak, M., Polychronakos, C., Tyler-Smith, C. & Kehrer-Sawatzki, H. 2013. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Human Genetics* **132:** 1077-1130.

Cooper, G. M. & Shendure, J. 2011. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nature Reviews.Genetics* **12:** 628-640.

Cottoni, E., Masia, I. M., Masala, M. V., Mulargia, M. & Contu, L. 1996. Familial Kaposi's sarcoma: case reports and review of the literature. *Acta Dermato-Venereologica* **76:** 59-61.

Cramer, S. F. & Patel, A. 1990. The frequency of uterine leiomyomas. *American Journal of Clinical Pathology* **94:** 435-438.

Czene, K., Lichtenstein, P. & Hemminki, K. 2002. Environmental and heritable causes of cancer among 9.6 million individuals in the Swedish Family-Cancer Database. *International Journal of Cancer* **99:** 260-266.

Dai, M., Wang, P., Boyd, A. D., Kostov, G., Athey, B., Jones, E. G., Bunney, W. E., Myers, R. M., Speed, T. P., Akil, H., Watson, S. J. & Meng, F. 2005. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Research* **33:** e175.

Day-Williams, A. G. & Zeggini, E. 2011. The effect of next-generation sequencing technology on complex trait research. *European Journal of Clinical Investigation* **41:** 561-567.

de Ligt, J., Willemsen, M. H., van Bon, B. W., Kleefstra, T., Yntema, H. G., Kroes, T., Vulto-van Silfhout, A. T., Koolen, D. A., de Vries, P., Gilissen, C., del Rosario, M., Hoischen, A., Scheffer, H., de Vries, B. B., Brunner, H. G., Veltman, J. A. & Vissers, L. E. 2012. Diagnostic exome sequencing in persons with severe intellectual disability. *The New England Journal of Medicine* **367:** 1921-1929.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D. & Daly, M. J. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* **43:** 491-498.

DiGiovanna, J. J. & Safai, B. 1981. Kaposi's sarcoma. Retrospective study of 90 cases with particular emphasis on the familial occurrence, ethnic background and prevalence of other diseases. *The American Journal of Medicine* **71:** 779-783.

Drmanac, R., Sparks, A. B., Callow, M. J., Halpern, A. L., Burns, N. L., Kermani, B. G., Carnevali, P., Nazarenko, I., Nilsen, G. B., Yeung, G., Dahl, F., Fernandez, A., Staker, B., Pant, K. P., Baccash, J., Borcherding, A. P., Brownley, A., Cedeno, R., Chen, L., Chernikoff, D., Cheung, A., Chirita, R., Curson, B., Ebert, J. C., Hacker, C. R., Hartlage, R., Hauser, B., Huang, S., Jiang, Y., Karpinchyk, V., Koenig, M., Kong, C., Landers, T., Le, C., Liu, J., McBride, C. E., Morenzoni, M., Morey, R. E., Mutch, K., Perazich, H., Perry, K., Peters, B. A., Peterson, J., Pethiyagoda, C. L., Pothuraju, K., Richter, C., Rosenbaum, A. M., Roy, S., Shafto, J., Sharanhovich, U., Shannon, K. W., Sheppy, C. G., Sun, M., Thakuria, J. V., Tran, A., Vu, D., Zaranek, A. W., Wu, X., Drmanac, S., Oliphant, A. R., Banyai, W. C., Martin, B., Ballinger, D. G., Church, G. M. & Reid, C. A. 2010. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327:** 78-81.

Elston, R. C. & Stewart, J. 1971. A general model for the genetic analysis of pedigree data. *Human Heredity* **21:** 523-542.

Eronen, M., Kajantie, E., Boldt, T., Pitkänen, O. & Aittomäki, K. 2004. Right atrial isomerism in four siblings. *Pediatric Cardiology* **25:** 141-144.

Eronen, M. P., Aittomäki, K. A., Kajantie, E. O., Sairanen, H. I. & Pesonen, E. J. 2013. The outcome of patients with right atrial isomerism is poor. *Pediatric Cardiology* **34:** 302-307.

Firth, H. V., Richards, S. M., Bevan, A. P., Clayton, S., Corpas, M., Rajan, D., Van Vooren, S., Moreau, Y., Pettett, R. M. & Carter, N. P. 2009. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *American Journal of Human Genetics* **84:** 524-533.

Flake, G. P., Andersen, J. & Dixon, D. 2003. Etiology and pathogenesis of uterine leiomyomas: a review. *Environmental Health Perspectives* **111:** 1037-1054.

Forbes, S. A., Bindal, N., Bamford, S., Cole, C., Kok, C. Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A., Teague, J. W., Campbell, P. J., Stratton, M. R. & Futreal, P. A. 2011. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Research* **39:** D945-50.

Freedom, R. M., Jaeggi, E. T., Lim, J. S. & Anderson, R. H. 2005. Hearts with isomerism of the right atrial appendages - one of the worst forms of disease in 2005. *Cardiology in the young* **15:** 554-567.

Gebbia, M., Ferrero, G. B., Pilia, G., Bassi, M. T., Aylsworth, A., Penman-Splitt, M., Bird, L. M., Bamforth, J. S., Burn, J., Schlessinger, D., Nelson, D. L. & Casey, B. 1997. X-linked situs abnormalities result from mutations in ZIC3. *Nature Genetics* **17:** 305-308.

Goldgar, D. E., Easton, D. F., Cannon-Albright, L. A. & Skolnick, M. H. 1994. Systematic population-based assessment of cancer risk in first-degree relatives of cancer probands. *Journal of the National Cancer Institute* **86:** 1600-1608.

Gunderson, K. L., Steemers, F. J., Lee, G., Mendoza, L. G. & Chee, M. S. 2005. A genome-wide scalable SNP genotyping assay using microarray technology. *Nature Genetics* **37:** 549-554.

Guo, S. W. & Thompson, E. A. 1992. A Monte Carlo method for combined segregation and linkage analysis. *American Journal of Human Genetics* **51:** 1111-1126.

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. 2005. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research* **33:** D514-7.

Hanahan, D. & Weinberg, R. A. 2011. Hallmarks of cancer: the next generation. *Cell* **144:** 646-674.

Hansen, M. F. & Cavenee, W. K. 1987. Genetics of cancer predisposition. *Cancer Research* **47:** 5518-5527.

Heidet, L., Boye, E., Cai, Y., Sado, Y., Zhang, X., Flejou, J. F., Fekete, F., Ninomiya, Y., Gubler, M. C. & Antignac, C. 1998. Somatic deletion of the 5' ends of both the COL4A5 and COL4A6 genes in a sporadic leiomyoma of the esophagus. *The American Journal of Pathology* **152:** 673-678.

Hemminki, K. & Li, X. 2001. Familial carcinoid tumors and subsequent cancers: a nation-wide epidemiologic study from Sweden. *International Journal of Cancer* **94:** 444-448.

Heutink, P. & Oostra, B. A. 2002. Gene finding in genetically isolated populations. *Human Molecular Genetics* **11:** 2507-2515.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S. & Manolio, T. A. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America* **106:** 9362-9367.

Hiripi, E., Bermejo, J. L., Sundquist, J. & Hemminki, K. 2009. Familial gastrointestinal carcinoid tumours and associated cancers. *Annals of Oncology : Official Journal of the European Society for Medical Oncology / ESMO* **20:** 950-954.

Hodge, J. C., Kim, T. M., Dreyfuss, J. M., Somasundaram, P., Christacos, N. C., Rousselle, M., Quade, B. J., Park, P. J., Stewart, E. A. & Morton, C. C. 2012. Expression profiling of uterine leiomyomata cytogenetic subgroups reveals distinct signatures in matched myometrium: transcriptional profilingof the t(12;14) and evidence in support of predisposing genetic heterogeneity. *Human Molecular Genetics* **21:** 2312-2329.

Holder, D., Raubertas, R.F., Pikounis, V.B., Svetnik, V. & Soper, K. 2001. Statistical analysis of high density oligonucleotide arrays: A SAFER approach. In: *GeneLogic Workshop on Low Level Analysis of Affymetrix GeneChip Data.*

Horsthemke, B. & Wagstaff, J. 2008. Mechanisms of imprinting of the Prader-Willi/Angelman region. *American Journal of Medical Genetics.Part A* **146A:** 2041-2052.

Huck, J. H., Verhoeven, N. M., Struys, E. A., Salomons, G. S., Jakobs, C. & van der Knaap, M. S. 2004. Ribose-5-phosphate isomerase deficiency: new inborn error in the pentose phosphate pathway associated with a slowly progressive leukoencephalopathy. *American Journal of Human Genetics* **74:** 745-751.

Hyvarinen, J., Parikka, M., Sormunen, R., Ramet, M., Tryggvason, K., Kivirikko, K. I., Myllyharju, J. & Koivunen, P. 2010. Deficiency of a transmembrane prolyl 4-hydroxylase in the zebrafish leads to basement membrane defects and compromised kidney function. *The Journal of Biological Chemistry* **285:** 42023-42032.

IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. 2012. Biological agents. Volume 100 B. A review of human carcinogens. *IARC monographs on the evaluation of carcinogenic risks to humans / World Health Organization, International Agency for Research on Cancer* **100:** 1-441.

Ingraham, S. E., Lynch, R. A., Kathiresan, S., Buckler, A. J. & Menon, A. G. 1999. hREC2, a RAD51-like gene, is disrupted by t(12;14) (q15;q24.1) in a uterine leiomyoma. *Cancer Genetics and Cytogenetics* **115:** 56-61.

International HapMap 3 Consortium, Altshuler, D. M., Gibbs, R. A., Peltonen, L., Altshuler, D. M., Gibbs, R. A., Peltonen, L., Dermitzakis, E., Schaffner, S. F., Yu, F., Peltonen, L., Dermitzakis, E., Bonnen, P. E., Altshuler, D. M., Gibbs, R. A., de Bakker, P. I., Deloukas, P., Gabriel, S. B., Gwilliam, R., Hunt, S., Inouye, M., Jia, X., Palotie, A., Parkin, M., Whittaker, P., Yu, F., Chang, K., Hawes, A., Lewis, L. R., Ren, Y., Wheeler, D., Gibbs, R. A., Muzny, D. M., Barnes, C., Darvishi, K., Hurles, M., Korn, J. M., Kristiansson, K., Lee, C., McCarrol, S. A., Nemesh, J., Dermitzakis, E., Keinan, A., Montgomery, S. B., Pollack, S., Price, A. L., Soranzo, N., Bonnen, P. E., Gibbs, R. A., Gonzaga-Jauregui, C., Keinan, A., Price, A. L., Yu, F., Anttila, V., Brodeur, W., Daly, M. J., Leslie, S., McVean, G., Moutsianas, L., Nguyen, H., Schaffner, S. F., Zhang, Q., Ghori, M. J., McGinnis, R., McLaren, W., Pollack, S., Price, A. L., Schaffner, S. F., Takeuchi, F., Grossman, S. R., Shlyakhter, I., Hostetter, E. B., Sabeti, P. C., Adebamowo, C. A., Foster, M. W., Gordon, D. R., Licinio, J., Manca, M. C., Marshall, P. A., Matsuda, I., Ngare, D., Wang, V. O., Reddy, D., Rotimi, C. N., Royal, C. D., Sharp, R. R., Zeng, C., Brooks, L. D. & McEwen, J. E. 2010. Integrating common and rare genetic variation in diverse human populations. *Nature* **467:** 52-58.

International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* **437:** 1299-1320.

International HapMap Consortium. 2003. The International HapMap Project. *Nature* **426:** 789-796.

International HapMap Consortium, Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., Boudreau, A., Hardenbol, P., Leal, S. M., Pasternak, S., Wheeler, D. A., Willis, T. D., Yu, F., Yang, H., Zeng, C., Gao, Y., Hu, H., Hu, W., Li, C., Lin, W., Liu, S., Pan, H., Tang, X., Wang, J., Wang, W., Yu, J., Zhang, B., Zhang, Q., Zhao, H., Zhao, H., Zhou, J., Gabriel, S. B., Barry, R., Blumenstiel, B., Camargo, A., Defelice, M., Faggart, M., Goyette, M., Gupta, S., Moore, J., Nguyen, H., Onofrio, R. C., Parkin, M., Roy, J., Stahl, E., Winchester, E., Ziaugra, L., Altshuler, D., Shen, Y., Yao, Z., Huang, W., Chu, X., He, Y., Jin, L., Liu, Y., Shen, Y., Sun, W., Wang, H., Wang, Y., Wang, Y., Xiong, X., Xu, L., Waye, M. M., Tsui, S. K., Xue, H., Wong, J. T., Galver, L. M., Fan, J. B., Gunderson, K., Murray, S. S., Oliphant, A. R., Chee, M. S., Montpetit, A., Chagnon, F., Ferretti, V., Leboeuf, M., Olivier, J. F., Phillips, M. S., Roumy, S., Sallee, C., Verner, A., Hudson, T. J., Kwok, P. Y., Cai, D., Koboldt, D. C., Miller, R. D., Pawlikowska, L., Taillon-Miller, P., Xiao, M., Tsui, L. C., Mak, W., Song, Y. Q., Tam, P. K., Nakamura, Y., Kawaguchi, T., Kitamoto, T., Morizono, T., Nagashima, A., Ohnishi, Y., Sekine, A., Tanaka, T., Tsunoda, T., Deloukas, P., Bird, C. P., Delgado, M., Dermitzakis, E. T., Gwilliam, R., Hunt, S., Morrison, J., Powell, D., Stranger, B. E., Whittaker, P., Bentley, D. R., Daly, M. J., de Bakker, P. I., Barrett, J., Chretien, Y. R., Maller, J., McCarroll, S., Patterson, N., Pe'er, I., Price, A., Purcell, S., Richter, D. J., Sabeti, P., Saxena, R., Schaffner, S. F., Sham, P. C., Varilly, P., Altshuler, D., Stein, L. D., Krishnan, L., Smith, A. V., Tello-Ruiz, M. K., Thorisson, G. A., Chakravarti, A., Chen, P. E., Cutler, D. J., Kashuk, C. S., Lin, S., Abecasis, G. R., Guan, W., Li, Y., Munro, H. M., Qin, Z. S., Thomas, D. J., McVean, G., Auton, A., Bottolo, L., Cardin, N., Eyheramendy, S., Freeman, C., Marchini, J., Myers, S., Spencer, C., Stephens, M., Donnelly, P., Cardon, L. R., Clarke, G., Evans, D. M., Morris, A. P., Weir, B. S., Tsunoda, T., Mullikin, J. C., Sherry, S. T., Feolo, M., Skol, A., Zhang, H., Zeng, C., Zhao, H., Matsuda, I., Fukushima, Y., Macer, D. R., Suda, E., Rotimi, C. N., Adebamowo, C. A., Ajayi, I., Aniagwu, T., Marshall, P. A., Nkwodimmah, C., Royal, C. D., Leppert, M. F., Dixon, M., Peiffer, A., Qiu, R., Kent, A., Kato, K., Niikawa, N., Adewole, I. F., Knoppers, B. M., Foster, M. W., Clayton, E. W., Watkin, J., Gibbs, R. A., Belmont, J. W., Muzny, D., Nazareth, L., Sodergren, E., Weinstock, G. M., Wheeler, D. A., Yakub,

I., Gabriel, S. B., Onofrio, R. C., Richter, D. J., Ziaugra, L., Birren, B. W., Daly, M. J., Altshuler, D., Wilson, R. K., Fulton, L. L., Rogers, J., Burton, J., Carter, N. P., Clee, C. M., Griffiths, M., Jones, M. C., McLay, K., Plumb, R. W., Ross, M. T., Sims, S. K., Willey, D. L., Chen, Z., Han, H., Kang, L., Godbout, M., Wallenburg, J. C., L'Archeveque, P., Bellemare, G., Saeki, K., Wang, H., An, D., Fu, H., Li, Q., Wang, Z., Wang, R., Holden, A. L., Brooks, L. D., McEwen, J. E., Guyer, M. S., Wang, V. O., Peterson, J. L., Shi, M., Spiegel, J., Sung, L. M., Zacharia, L. F., Collins, F. S., Kennedy, K., Jamieson, R. & Stewart, J. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449:** 851-861.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431:** 931-945.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U. & Speed, T. P. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4:** 249-264.

Ivemark, B. I. 1955. Implications of agenesis of the spleen on the pathogenesis of conotruncus anomalies in childhood; an analysis of the heart malformations in the splenic agenesis syndrome, with fourteen new cases. *Acta paediatrica.Supplementum* **44:** 7-110.

Jakkula, E., Rehnstrom, K., Varilo, T., Pietilainen, O. P., Paunio, T., Pedersen, N. L., deFaire, U., Jarvelin, M. R., Saharinen, J., Freimer, N., Ripatti, S., Purcell, S., Collins, A., Daly, M. J., Palotie, A. & Peltonen, L. 2008. The genome-wide patterns of variation expose significant substructure in a founder population. *American Journal of Human Genetics* **83:** 787-794.

Joensuu, T., Kuronen, M., Alakurtti, K., Tegelberg, S., Hakala, P., Aalto, A., Huopaniemi, L., Aula, N., Michellucci, R., Eriksson, K. & Lehesjoki, A. E. 2007. Cystatin B: mutation detection, alternative splicing and expression in progressive myclonus epilepsy of Unverricht-Lundborg type (EPM1) patients. *European Journal of Human Genetics* **15:** 185-193.

Jones, M. J. & Jallepalli, P. V. 2012. Chromothripsis: chromosomes in crisis. *Developmental Cell* **23:** 908-917.

Kaposi, M. 1872. Idiopathisches multiples Pigmentsarkom der Haut. *Arch. Dermatol. Syph.* **4**: 265–273

Karkera, J. D., Lee, J. S., Roessler, E., Banerjee-Basu, S., Ouspenskaia, M. V., Mez, J., Goldmuntz, E., Bowers, P., Towbin, J., Belmont, J. W., Baxevanis, A. D., Schier, A. F. & Muenke, M. 2007. Loss-of-function mutations in growth differentiation factor-1 (GDF1) are associated with congenital heart defects in humans. *American Journal of Human Genetics* **81:** 987-994.

Kato, R., Yamada, Y. & Niikawa, N. 1996. De novo balanced translocation (6;18)(q21;q21.3 or q22) [corrected] in a patient with heterotaxia. *American Journal of Medical Genetics* **66:** 184-186.

Kennedy, M. P., Omran, H., Leigh, M. W., Dell, S., Morgan, L., Molina, P. L., Robinson, B. V., Minnix, S. L., Olbrich, H., Severin, T., Ahrens, P., Lange, L., Morillas, H. N., Noone, P. G., Zariwala, M. A. & Knowles, M. R. 2007. Congenital heart disease and other heterotaxic defects in a large cohort of patients with primary ciliary dyskinesia. *Circulation* **115:** 2814-2821.

Kharazmi, E., Pukkala, E., Sundquist, K. & Hemminki, K. 2013. Familial risk of small intestinal carcinoid and adenocarcinoma. *Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association* **11:** 944-949.

Khoshnoodi, J., Pedchenko, V. & Hudson, B. G. 2008. Mammalian collagen IV. *Microscopy Research and Technique* **71:** 357-370.

Khoury, M. J., Beaty, T. H. & Cohen, B. H. 1993. Fundamentals of genetic epidemiology. Oxford University Press, USA,

Khoury, M. J., Gwinn, M., Clyne, M. & Yu, W. 2011. Genetic epidemiology with a capital E, ten years after. *Genetic Epidemiology* **35:** 845-852.

King, M., Lee, G. M., Spinner, N. B., Thomson, G. & Wrensch, M. R. 1984. Genetic epidemiology. *Annual Review of Public Health* **5:** 1-52.

Kitts, A., Phan, L., Ward, M. & Church, D. (Editors). 2013. The database of short genetic variation (dbSNP) 2013 jun 30. in: The NCBI handbook [internet]. 2nd edition edition. Bethesda (MD): National Center for Biotechnology Information (US), http://www.ncbi.nlm.nih.gov/books/NBK174586/.

Kiuru, M., Launonen, V., Hietala, M., Aittomäki, K., Vierimaa, O., Salovaara, R., Arola, J., Pukkala, E., Sistonen, P., Herva, R. & Aaltonen, L. A. 2001. Familial cutaneous leiomyomatosis is a two-hit condition associated with renal cell cancer of characteristic histopathology. *The American Journal of Pathology* **159:** 825-829.

Knudson, A. G.,Jr. 1971. Mutation and cancer: statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences of the United States of America* **68:** 820-823.

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., Miller, C. A., Mardis, E. R., Ding, L. & Wilson, R. K. 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research* **22:** 568-576.

Koivunen, P., Tiainen, P., Hyvarinen, J., Williams, K. E., Sormunen, R., Klaus, S. J., Kivirikko, K. I. & Myllyharju, J. 2007. An endoplasmic reticulum transmembrane prolyl 4-hydroxylase is induced by hypoxia and acts on hypoxia-inducible factor alpha. *The Journal of Biological Chemistry* **282:** 30544-30552.

Kosaki, K., Bassi, M. T., Kosaki, R., Lewin, M., Belmont, J., Schauer, G. & Casey, B. 1999a. Characterization and mutation analysis of human LEFTY A and LEFTY B, homologues of murine genes implicated in left-right axis development. *American Journal of Human Genetics* **64:** 712-721.

Kosaki, R., Gebbia, M., Kosaki, K., Lewin, M., Bowers, P., Towbin, J. A. & Casey, B. 1999b. Left-right axis malformations associated with mutations in ACVR2B, the gene for human activin receptor type IIB. *American Journal of Medical Genetics* **82:** 70-76.

Kruglyak, L., Daly, M. J., Reeve-Daly, M. P. & Lander, E. S. 1996. Parametric and nonparametric linkage analysis: a unified multipoint approach. *American Journal of Human Genetics* **58:** 1347-1363.

Lander, E. S. & Green, P. 1987. Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences of the United States of America* **84:** 2363-2367.

Launonen, V., Vierimaa, O., Kiuru, M., Isola, J., Roth, S., Pukkala, E., Sistonen, P., Herva, R. & Aaltonen, L. A. 2001. Inherited susceptibility to uterine leiomyomas and renal cell cancer. *Proceedings of the National Academy of Sciences of the United States of America* **98:** 3387-3392.

Lehtonen, R., Kiuru, M., Vanharanta, S., Sjöberg, J., Aaltonen, L. M., Aittomäki, K., Arola, J., Bützow, R., Eng, C., Husgafvel-Pursiainen, K., Isola, J., Järvinen, H., Koivisto, P., Mecklin, J. P., Peltomäki, P., Salovaara, R., Wasenius, V. M., Karhu, A., Launonen, V., Nupponen, N. N. & Aaltonen, L. A. 2004. Biallelic inactivation of fumarate hydratase (FH) occurs in nonsyndromic uterine leiomyomas but is rare in other tumors. *The American Journal of Pathology* **164:** 17-22.

Leibsohn, S., d'Ablaing, G., Mishell, D. R.,Jr & Schlaerth, J. B. 1990. Leiomyosarcoma in a series of hysterectomies performed for presumed uterine leiomyomas. *American Journal of Obstetrics and Gynecology* **162:** 968-74; discussion 974-6.

Levin, M. 2005. Left-right asymmetry in embryonic development: a comprehensive review. *Mechanisms of Development* **122:** 3-25.

Li, H. & Durbin, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25:** 1754-1760.

Li, H., Ruan, J. & Durbin, R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research* **18:** 1851-1858.

Lichtenstein, P., Holm, N. V., Verkasalo, P. K., Iliadou, A., Kaprio, J., Koskenvuo, M., Pukkala, E., Skytthe, A. & Hemminki, K. 2000. Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. *The New England Journal of Medicine* **343:** 78-85.

Ligon, A. H. & Morton, C. C. 2000. Genetics of uterine leiomyomata. *Genes, Chromosomes & Cancer* **28:** 235-245.

Ligon, A. H., Scott, I. C., Takahara, K., Greenspan, D. S. & Morton, C. C. 2002. PCOLCE deletion and expression analyses in uterine leiomyomata. *Cancer Genetics and Cytogenetics* **137:** 133-137.

Lim, J. S., McCrindle, B. W., Smallhorn, J. F., Golding, F., Caldarone, C. A., Taketazu, M. & Jaeggi, E. T. 2005. Clinical features, management, and outcome of children with fetal and postnatal diagnoses of isomerism syndromes. *Circulation* **112:** 2454-2461.

Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H. & Brown, E. L. 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology* **14:** 1675-1680.

Luoto, R., Kaprio, J., Rutanen, E. M., Taipale, P., Perola, M. & Koskenvuo, M. 2000. Heritability and risk factors of uterine fibroids--the Finnish Twin Cohort study. *Maturitas* **37:** 15-26.

MacArthur, D. G., Manolio, T. A., Dimmock, D. P., Rehm, H. L., Shendure, J., Abecasis, G. R., Adams, D. R., Altman, R. B., Antonarakis, S. E., Ashley, E. A., Barrett, J. C., Biesecker, L. G., Conrad, D. F., Cooper, G. M., Cox, N. J., Daly, M. J., Gerstein, M. B., Goldstein, D. B., Hirschhorn, J. N., Leal, S. M., Pennacchio, L. A., Stamatoyannopoulos, J. A., Sunyaev, S. R., Valle, D., Voight, B. F., Winckler, W. & Gunter, C. 2014. Guidelines for investigating causality of sequence variants in human disease. *Nature* **508:** 469-476.

Macdonald, J. R., Ziman, R., Yuen, R. K., Feuk, L. & Scherer, S. W. 2014. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Research* **42:** D986-992.

Mäkinen, N., Heinonen, H. R., Moore, S., Tomlinson, I. P., van der Spuy, Z. M. & Aaltonen, L. A. 2011a. MED12 exon 2 mutations are common in uterine leiomyomas from South African patients. *Oncotarget* **2:** 966-969.

Mäkinen, N., Mehine, M., Tolvanen, J., Kaasinen, E., Li, Y., Lehtonen, H. J., Gentile, M., Yan, J., Enge, M., Taipale, M., Aavikko, M., Katainen, R., Virolainen, E., Bohling, T., Koski, T. A., Launonen, V., Sjöberg, J., Taipale, J., Vahteristo, P. & Aaltonen, L. A. 2011b. MED12, the mediator complex subunit 12 gene, is mutated at high frequency in uterine leiomyomas. *Science* **334:** 252-255.

Mancuso, R., Brambilla, L., Agostini, S., Biffi, R., Hernis, A., Guerini, F. R., Agliardi, C., Tourlaki, A., Bellinvia, M. & Clerici, M. 2011. Intrafamiliar transmission of Kaposi's sarcoma-associated herpesvirus and seronegative infection in family members of classic Kaposi's sarcoma patients. *The Journal of General Virology* **92:** 744-751.

Manolio, T. A. 2010. Genomewide association studies and assessment of the risk of disease. *The New England Journal of Medicine* **363:** 166-176.

Markowski, D. N., Bartnitzke, S., Loning, T., Drieschner, N., Helmke, B. M. & Bullerdiek, J. 2012. MED12 mutations in uterine fibroids--their relationship to cytogenetic subgroups. *International Journal of Cancer* **131:** 1528-1536.

Marshall, L. M., Spiegelman, D., Barbieri, R. L., Goldman, M. B., Manson, J. E., Colditz, G. A., Willett, W. C. & Hunter, D. J. 1997. Variation in the incidence of uterine leiomyoma among premenopausal women by age and race. *Obstetrics and Gynecology* **90:** 967-973.

Mayr, C., Hemann, M. T. & Bartel, D. P. 2007. Disrupting the pairing between let-7 and Hmga2 enhances oncogenic transformation. *Science* **315:** 1576-1579.

Mbulaiteye, S. M. & Engels, E. A. 2006. Kaposi's sarcoma risk among transplant recipients in the United States (1993-2003). *International Journal of Cancer* **119:** 2685-2691.

McEvoy, J., Nagahawatte, P., Finkelstein, D., Richards-Yutz, J., Valentine, M., Ma, J., Mullighan, C., Song, G., Chen, X., Wilson, M., Brennan, R., Pounds, S., Becksfort, J., Huether, R., Lu, C., Fulton, R. S., Fulton, L. L., Hong, X., Dooling, D. J., Ochoa, K., Mardis, E. R., Wilson, R. K., Easton, J., Zhang, J., Downing, J. R., Ganguly, A. & Dyer, M. A. 2014. RB1 gene inactivation by chromothripsis in human retinoblastoma. *Oncotarget* **5**:438-450

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. & DePristo, M. A. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20:** 1297-1303.

McNerney, M. E., Brown, C. D., Wang, X., Bartom, E. T., Karmakar, S., Bandlamudi, C., Yu, S., Ko, J., Sandall, B. P., Stricker, T., Anastasi, J., Grossman, R. L., Cunningham, J. M., Le Beau, M. M. & White, K. P. 2013. CUX1 is a haploinsufficient tumor suppressor gene on chromosome 7 frequently inactivated in acute myeloid leukemia. *Blood* **121:** 975-983.

Mesri, E. A., Cesarman, E. & Boshoff, C. 2010. Kaposi's sarcoma and its associated herpesvirus. *Nature Reviews.Cancer* **10:** 707-719.

Metzker, M. L. 2010. Sequencing technologies - the next generation. *Nature Reviews.Genetics* **11:** 31-46.

Mills, R. E., Walter, K., Stewart, C., Handsaker, R. E., Chen, K., Alkan, C., Abyzov, A., Yoon, S. C., Ye, K., Cheetham, R. K., Chinwalla, A., Conrad, D. F., Fu, Y., Grubert, F., Hajirasouliha, I., Hormozdiari, F., Iakoucheva, L. M., Iqbal, Z., Kang, S., Kidd, J. M., Konkel, M. K., Korn, J., Khurana, E., Kural, D., Lam, H. Y., Leng, J., Li, R., Li, Y., Lin, C. Y., Luo, R., Mu, X. J., Nemesh, J., Peckham, H. E., Rausch, T., Scally, A., Shi, X., Stromberg, M. P., Stutz, A. M., Urban, A. E., Walker, J. A., Wu, J., Zhang, Y., Zhang, Z. D., Batzer, M. A., Ding, L., Marth, G. T., McVean, G., Sebat, J., Snyder, M., Wang, J., Ye, K., Eichler, E. E., Gerstein, M. B., Hurles, M. E., Lee, C., McCarroll, S. A., Korbel, J. O. & 1000 Genomes Project. 2011. Mapping copy number variation by population-scale genome sequencing. *Nature* **470:** 59-65.

Milojevic, T., Reiterer, V., Stefan, E., Korkhov, V. M., Dorostkar, M. M., Ducza, E., Ogris, E., Boehm, S., Freissmuth, M. & Nanoff, C. 2006. The ubiquitin-specific protease Usp4 regulates the cell surface level of the A2A receptor. *Molecular Pharmacology* **69:** 1083-1094.

Mitschke, L., Parthier, C., Schroder-Tittmann, K., Coy, J., Ludtke, S. & Tittmann, K. 2010. The crystal structure of human transketolase and new insights into its mode of action. *The Journal of Biological Chemistry* **285:** 31559-31570.

Mohapatra, B., Casey, B., Li, H., Ho-Dawson, T., Smith, L., Fernbach, S. D., Molinari, L., Niesh, S. R., Jefferies, J. L., Craigen, W. J., Towbin, J. A., Belmont, J. W. & Ware, S. M. 2009. Identification and functional characterization of NODAL rare variants in heterotaxy and isolated cardiovascular malformations. *Human Molecular Genetics* **18:** 861-871.

Mokry, M., Feitsma, H., Nijman, I. J., de Bruijn, E., van der Zaag, P. J., Guryev, V. & Cuppen, E. 2010. Accurate SNP and mutation detection by targeted custom microarray-based genomic enrichment of short-fragment sequencing libraries. *Nucleic Acids Research* **38:** e116.

Morton, N. E. 1955. Sequential tests for the detection of linkage. *American Journal of Human Genetics* **7:** 277-318.

Mukhopadhyay, N., Almasy, L., Schroeder, M., Mulvihill, W. P. & Weeks, D. E. 2005. Mega2: data-handling for facilitating genetic linkage and association analyses. *Bioinformatics* **21:** 2556-2557.

Musante, L. & Ropers, H. H. 2014. Genetics of recessive cognitive disorders. *Trends in Genetics : TIG* **30:** 32-39.

Najmabadi, H., Hu, H., Garshasbi, M., Zemojtel, T., Abedini, S. S., Chen, W., Hosseini, M., Behjati, F., Haas, S., Jamali, P., Zecha, A., Mohseni, M., Puttmann, L., Vahid, L. N., Jensen, C., Moheb, L. A., Bienek, M., Larti, F., Mueller, I., Weissmann, R., Darvish, H., Wrogemann, K., Hadavi, V., Lipkowitz, B., Esmaeeli-Nieh, S., Wieczorek, D., Kariminejad, R., Firouzabadi, S. G., Cohen, M., Fattahi, Z., Rost, I., Mojahedi, F., Hertzberg, C., Dehghan, A., Rajab, A., Banavandi, M. J., Hoffer, J., Falah, M., Musante, L., Kalscheuer, V., Ullmann, R., Kuss, A. W., Tzschach, A., Kahrizi, K. & Ropers, H. H. 2011. Deep sequencing reveals 50 novel genes for recessive cognitive disorders. *Nature* **478:** 57-63.

Nibert, M. & Heim, S. 1990. Uterine leiomyoma cytogenetics. *Genes, Chromosomes & Cancer* **2:** 3-13.

Nilbert, M., Heim, S., Mandahl, N., Floderus, U. M., Willen, H. & Mitelman, F. 1990. Characteristic chromosome abnormalities, including rearrangements of 6p, del(7q), +12, and t(12;14), in 44 uterine leiomyomas. *Human Genetics* **85:** 605-611.

Nonaka, S., Tanaka, Y., Okada, Y., Takeda, S., Harada, A., Kanai, Y., Kido, M. & Hirokawa, N. 1998. Randomization of left-right asymmetry due to loss of nodal cilia generating leftward flow of extraembryonic fluid in mice lacking KIF3B motor protein. *Cell* **95:** 829-837.

Noone, P. G., Leigh, M. W., Sannuti, A., Minnix, S. L., Carson, J. L., Hazucha, M., Zariwala, M. A. & Knowles, M. R. 2004. Primary ciliary dyskinesia: diagnostic and phenotypic features. *American Journal of Respiratory and Critical Care Medicine* **169:** 459-467.

Northcott, P. A., Shih, D. J., Peacock, J., Garzia, L., Morrissy, A. S., Zichner, T., Stutz, A. M., Korshunov, A., Reimand, J., Schumacher, S. E., Beroukhim, R., Ellison, D. W., Marshall, C. R., Lionel, A. C., Mack, S., Dubuc, A., Yao, Y., Ramaswamy, V., Luu, B., Rolider, A., Cavalli, F. M., Wang, X., Remke, M., Wu, X., Chiu, R. Y., Chu, A., Chuah, E., Corbett, R. D., Hoad, G. R., Jackman, S. D., Li, Y., Lo, A., Mungall, K. L., Nip, K. M., Qian, J. Q., Raymond, A. G., Thiessen, N. T., Varhol, R. J., Birol, I., Moore, R. A., Mungall, A. J., Holt, R., Kawauchi, D., Roussel, M. F., Kool, M., Jones, D. T., Witt, H., Fernandez-L, A., Kenney, A. M., Wechsler-Reya, R. J., Dirks, P., Aviv, T., Grajkowska, W. A., Perek-Polnik, M., Haberler, C. C., Delattre, O., Reynaud, S. S., Doz, F. F., Pernet-Fattet, S. S., Cho, B. K., Kim, S. K., Wang, K. C., Scheurlen, W., Eberhart, C. G., Fevre-Montange, M., Jouvet, A., Pollack, I. F., Fan, X., Muraszko, K. M., Gillespie, G. Y., Di Rocco, C., Massimi, L., Michiels, E. M., Kloosterhof, N. K., French, P. J., Kros, J. M., Olson, J. M., Ellenbogen, R. G., Zitterbart, K., Kren, L., Thompson, R. C., Cooper, M. K., Lach, B., McLendon, R. E., Bigner, D. D., Fontebasso, A., Albrecht, S., Jabado, N., Lindsey, J. C., Bailey, S., Gupta, N., Weiss, W. A., Bognar, L., Klekner, A., Van Meter, T. E., Kumabe, T., Tominaga, T., Elbabaa, S. K., Leonard, J. R., Rubin, J. B., Liau, L. M., Van Meir, E. G., Fouladi, M., Nakamura, H., Cinalli, G., Garami, M., Hauser, P., Saad, A. G., Iolascon, A., Jung, S., Carlotti, C. G., Vibhakar, R., Ra, Y. S., Robinson, S., Zollo, M., Faria, C. C., Chan, J. A., Levy, M. L., Sorensen, P. H., Meyerson, M., Pomeroy, S. L., Cho, Y. J., Bader, G. D., Tabori, U., Hawkins, C. E., Bouffet, E., Scherer, S. W., Rutka, J. T., Malkin, D., Clifford, S. C., Jones, S. J., Korbel, J. O., Pfister, S. M., Marra, M. A. & Taylor, M. D. 2012. Subgroup-specific structural variation across 1,000 medulloblastoma genomes. *Nature* **488:** 49-56.

Okada, Y., Nonaka, S., Tanaka, Y., Saijoh, Y., Hamada, H. & Hirokawa, N. 1999. Abnormal nodal flow precedes situs inversus in iv and inv mice. *Molecular Cell* **4:** 459-468.

Okumura, T., Utsuno, H., Kuroda, J., Gittenberger, E., Asami, T. & Matsuno, K. 2008. The development and evolution of left-right asymmetry in invertebrates: lessons from Drosophila and snails. *Developmental Dynamics : an official publication of the American Association of Anatomists* **237:** 3497-3515.

Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M. R., Zschocke, J. & Trajanoski, Z. 2014. A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in Bioinformatics* **15:** 256-278.

Pang, A. W., MacDonald, J. R., Pinto, D., Wei, J., Rafiq, M. A., Conrad, D. F., Park, H., Hurles, M. E., Lee, C., Venter, J. C., Kirkness, E. F., Levy, S., Feuk, L. & Scherer, S. W. 2010. Towards a comprehensive structural variation map of an individual human genome. *Genome Biology* **11:** R52-2010-11-5-r52. Epub 2010 May 19.

Parker, M., Mohankumar, K. M., Punchihewa, C., Weinlich, R., Dalton, J. D., Li, Y., Lee, R., Tatevossian, R. G., Phoenix, T. N., Thiruvenkatam, R., White, E., Tang, B., Orisme, W., Gupta, K., Rusch, M., Chen, X., Li, Y., Nagahawhatte, P., Hedlund, E., Finkelstein, D., Wu, G., Shurtleff, S., Easton, J., Boggs, K., Yergeau, D., Vadodaria, B., Mulder, H. L., Becksford, J., Gupta, P., Huether, R., Ma, J., Song, G., Gajjar, A., Merchant, T., Boop, F., Smith, A. A., Ding, L., Lu, C., Ochoa, K., Zhao, D., Fulton, R. S., Fulton, L. L., Mardis, E. R., Wilson, R. K., Downing, J. R., Green, D. R., Zhang, J., Ellison, D. W. & Gilbertson, R. J. 2014. C11orf95-RELA fusions drive oncogenic NF-kappaB signalling in ependymoma. *Nature* **506:** 451-455.

Peeters, H., Debeer, P., Groenen, P., Van Esch, H., Vanderlinden, G., Eyskens, B., Mertens, L., Gewillig, M., Van de Ven, W., Fryns, J. P. & Devriendt, K. 2001. Recurrent involvement of chromosomal region 6q21 in heterotaxy. *American Journal of Medical Genetics* **103:** 44-47.

Peiffer, D. A., Le, J. M., Steemers, F. J., Chang, W., Jenniges, T., Garcia, F., Haden, K., Li, J., Shaw, C. A., Belmont, J., Cheung, S. W., Shen, R. M., Barker, D. L. & Gunderson, K. L. 2006. High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Research* **16:** 1136-1148.

Peltonen, L., Jalanko, A. & Varilo, T. 1999a. Molecular genetics of the Finnish disease heritage. *Human Molecular Genetics* **8:** 1913-1923.

Peltonen, L., Jalanko, A. & Varilo, T. 1999b. Molecular genetics of the Finnish disease heritage. *Human Molecular Genetics* **8:** 1913-1923.

Perles, Z., Cinnamon, Y., Ta-Shma, A., Shaag, A., Einbinder, T., Rein, A. J. & Elpeleg, O. 2012. A human laterality disorder associated with recessive CCDC11 mutation. *Journal of Medical Genetics* **49:** 386-390.

Petrij, F., Giles, R. H., Dauwerse, H. G., Saris, J. J., Hennekam, R. C., Masuno, M., Tommerup, N., van Ommen, G. J., Goodman, R. H. & Peters, D. J. 1995. Rubinstein-Taybi syndrome caused by mutations in the transcriptional co-activator CBP. *Nature* **376:** 348-351.

Picard, C., Mellouli, F., Duprez, R., Chedeville, G., Neven, B., Fraitag, S., Delaunay, J., Le Deist, F., Fischer, A., Blanche, S., Bodemer, C., Gessain, A., Casanova, J. L. & Bejaoui, M. 2006. Kaposi's sarcoma in a child with Wiskott-Aldrich syndrome. *European Journal of Pediatrics* **165:** 453-457.

Pickrell, J. K., Gaffney, D. J., Gilad, Y. & Pritchard, J. K. 2011. False positive peaks in ChIP-seq and other sequencing-based functional assays caused by unannotated high copy number regions. *Bioinformatics* **27:** 2144-2146.

Pukkala, E. 2011. Biobanks and registers in epidemiologic research on cancer. *Methods in Molecular Biology* **675:** 127-164.

Quintana, D. G., Thome, K. C., Hou, Z. H., Ligon, A. H., Morton, C. C. & Dutta, A. 1998. ORC5L, a new member of the human origin recognition complex, is deleted in uterine leiomyomas and malignant myeloid diseases. *The Journal of Biological Chemistry* **273:** 27137-27145.

Rahman, N. 2014. Realizing the promise of cancer predisposition genes. *Nature* **505:** 302-308.

Rankin, C. T., Bunton, T., Lawler, A. M. & Lee, S. J. 2000. Regulation of left-right patterning in mice by growth/differentiation factor-1. *Nature Genetics* **24:** 262-265.

Rausch, T., Jones, D. T., Zapatka, M., Stutz, A. M., Zichner, T., Weischenfeldt, J., Jager, N., Remke, M., Shih, D., Northcott, P. A., Pfaff, E., Tica, J., Wang, Q., Massimi, L., Witt, H., Bender, S., Pleier, S., Cin, H., Hawkins, C., Beck, C., von Deimling, A., Hans, V., Brors, B., Eils, R., Scheurlen, W., Blake, J., Benes, V., Kulozik, A. E., Witt, O., Martin, D., Zhang, C., Porat, R., Merino, D. M., Wasserman, J., Jabado, N., Fontebasso, A., Bullinger, L., Rucker, F. G., Dohner, K., Dohner, H., Koster, J., Molenaar, J. J., Versteeg, R., Kool, M., Tabori, U., Malkin, D., Korshunov, A., Taylor, M. D., Lichter, P., Pfister, S. M. & Korbel, J. O. 2012. Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. *Cell* **148:** 59-71.

Raya, A. & Belmonte, J. C. 2006. Left-right asymmetry in the vertebrate embryo: from early information to higher-level integration. *Nature Reviews.Genetics* **7:** 283-293.

Rein, M. S. 2000. Advances in uterine leiomyoma research: the progesterone hypothesis. *Environmental Health Perspectives* **108 Suppl 5:** 791-793.

Ritchie, G. R., Dunham, I., Zeggini, E. & Flicek, P. 2014. Functional annotation of noncoding sequence variants. *Nature Methods* **11**:294-296

Ropers, H. H. 2010. Genetics of early onset cognitive impairment. *Annual Review of Genomics and Human Genetics* **11:** 161-187.

Rousseau, F., Rouillard, P., Morel, M. L., Khandjian, E. W. & Morgan, K. 1995. Prevalence of carriers of premutation-size alleles of the FMRI gene--and implications for the population genetics of the fragile X syndrome. *American Journal of Human Genetics* **57:** 1006-1018.

Rozen, S. & Skaletsky, H. 2000. Primer3 on the WWW for general users and for biologist programmers. *Methods in Molecular Biology* **132:** 365-386.

Safai, B., Johnson, K. G., Myskowski, P. L., Koziner, B., Yang, S. Y., Cunningham-Rundles, S., Godbold, J. H. & Dupont, B. 1985. The natural history of Kaposi's sarcoma in the acquired immunodeficiency syndrome. *Annals of Internal Medicine* **103:** 744-750.

Santos, S. I. 1999. Cancer epidemiology, principles and methods. *Cancer Epidemiology, Principles and Methods*

Schenk, G., Duggleby, R. G. & Nixon, P. F. 1998. Properties and functions of the thiamin diphosphate dependent enzyme transketolase. *The International Journal of Biochemistry & Cell Biology* **30:** 1297-1318.

Schoenmakers, E. F., Bunt, J., Hermers, L., Schepens, M., Merkx, G., Janssen, B., Kersten, M., Huys, E., Pauwels, P., Debiec-Rychter, M. & van Kessel, A. G. 2013. Identification of CUX1 as the recurrent chromosomal band 7q22 target gene in human uterine leiomyoma. *Genes, Chromosomes & Cancer* **52:** 11-23.

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Maner, S., Massa, H., Walker, M., Chi, M., Navin, N., Lucito, R., Healy, J., Hicks, J., Ye, K., Reiner, A., Gilliam, T. C., Trask, B., Patterson, N., Zetterberg, A. & Wigler, M. 2004. Large-scale copy number polymorphism in the human genome. *Science* **305:** 525-528.

Shen, H., Li, J., Zhang, J., Xu, C., Jiang, Y., Wu, Z., Zhao, F., Liao, L., Chen, J., Lin, Y., Tian, Q., Papasian, C. J. & Deng, H. W. 2013. Comprehensive characterization of human genome variation by high coverage whole-genome sequencing of forty four Caucasians. *PloS ONE* **8:** e59494.

Shen, M. M. 2007. Nodal signaling: developmental roles and regulation. *Development* **134:** 1023-1034.

Sobel, E. & Lange, K. 1996. Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *American Journal of Human Genetics* **58:** 1323-1337.

Steemers, F. J., Chang, W., Lee, G., Barker, D. L., Shen, R. & Gunderson, K. L. 2006. Whole-genome genotyping with the single-base extension assay. *Nature Methods* **3:** 31-33.

Steemers, F. J. & Gunderson, K. L. 2005. Illumina, Inc. *Pharmacogenomics* **6:** 777-782.

Stenson, P. D., Mort, M., Ball, E. V., Shaw, K., Phillips, A. D. & Cooper, D. N. 2014. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Human Genetics* **133**: 1-9.

Stephens, P. J., Greenman, C. D., Fu, B., Yang, F., Bignell, G. R., Mudie, L. J., Pleasance, E. D., Lau, K. W., Beare, D., Stebbings, L. A., McLaren, S., Lin, M. L., McBride, D. J., Varela, I., Nik-Zainal, S., Leroy, C., Jia, M., Menzies, A., Butler, A. P., Teague, J. W., Quail, M. A., Burton, J., Swerdlow, H., Carter, N. P., Morsberger, L. A., Iacobuzio-Donahue, C., Follows, G. A., Green, A. R., Flanagan, A. M., Stratton, M. R., Futreal, P. A. & Campbell, P. J. 2011. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144:** 27-40.

Stratton, M. R., Campbell, P. J. & Futreal, P. A. 2009. The cancer genome. *Nature* **458:** 719-724.

Sun, P. D. & Davies, D. R. 1995. The cystine-knot growth-factor superfamily. *Annual Review of Biophysics and Biomolecular Structure* **24:** 269-291.

Tanaka, C., Sakuma, R., Nakamura, T., Hamada, H. & Saijoh, Y. 2007. Long-range action of Nodal requires interaction with GDF1. *Genes & Development* **21:** 3272-3282.

Teppo, L., Pukkala, E. & Lehtonen, M. 1994. Data quality and quality control of a population-based cancer registry. Experience in Finland. *Acta Oncologica* **33:** 365-369.

Thielen, B. K., Barker, D. F., Nelson, R. D., Zhou, J., Kren, S. M. & Segal, Y. 2003. Deletion mapping in Alport syndrome and Alport syndrome-diffuse leiomyomatosis reveals potential mechanisms of visceral smooth muscle overgrowth. *Human Mutation* **22:** 419.

Tomita-Mitchell, A., Mahnke, D. K., Struble, C. A., Tuffnell, M. E., Stamm, K. D., Hidestrand, M., Harris, S. E., Goetsch, M. A., Simpson, P. M., Bick, D. P., Broeckel, U., Pelech, A. N., Tweddell, J. S. & Mitchell, M. E. 2012. Human gene copy number spectra analysis in congenital heart malformations. *Physiological Genomics* **44:** 518-541.

Tomlinson, I. P., Alam, N. A., Rowan, A. J., Barclay, E., Jaeger, E. E., Kelsell, D., Leigh, I., Gorman, P., Lamlum, H., Rahman, S., Roylance, R. R., Olpin, S., Bevan, S., Barker, K., Hearle, N., Houlston, R. S., Kiuru, M., Lehtonen, R., Karhu, A., Vilkki, S., Laiho, P., Eklund, C., Vierimaa, O., Aittomäki, K., Hietala, M., Sistonen, P., Paetau, A., Salovaara, R., Herva, R., Launonen, V., Aaltonen, L. A. & Multiple Leiomyoma Consortium. 2002. Germline mutations in FH predispose to dominantly inherited uterine fibroids, skin leiomyomata and papillary renal cell cancer. *Nature Genetics* **30:** 406-410.

Tommerup, N. 1993. Mendelian cytogenetics. Chromosome rearrangements associated with mendelian disorders. *Journal of Medical Genetics* **30:** 713-727.

Treangen, T. J. & Salzberg, S. L. 2011. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews.Genetics* **13:** 36-46.

Trivier, E., De Cesare, D., Jacquot, S., Pannetier, S., Zackai, E., Young, I., Mandel, J. L., Sassone-Corsi, P. & Hanauer, A. 1996. Mutations in the kinase Rsk-2 associated with Coffin-Lowry syndrome. *Nature* **384:** 567-570.

Trollmann, R. & Gassmann, M. 2009. The role of hypoxia-inducible transcription factors in the hypoxic neonatal brain. *Brain & Development* **31:** 503-509.

Tuominen, I., Heliovaara, E., Raitila, A., Rautiainen, M. R., Mehine, M., Katainen, R., Donner, I., Aittomäki, V., Lehtonen, H. J., Ahlsten, M., Kivipelto, L., Schalin-Jantti, C., Arola, J., Hautaniemi, S. & Karhu, A. 2014. AIP inactivation leads to pituitary tumorigenesis through defective Galpha-cAMP signaling. *Oncogene*. Epub 2014 March 24; doi: 10.1038/onc.2014.50.

Vadnais, C., Davoudi, S., Afshin, M., Harada, R., Dudley, R., Clermont, P. L., Drobetsky, E. & Nepveu, A. 2012. CUX1 transcription factor is required for optimal ATM/ATR-mediated responses to DNA damage. *Nucleic Acids Research* **40:** 4483-4495.

Vanharanta, S., Pollard, P. J., Lehtonen, H. J., Laiho, P., Sjöberg, J., Leminen, A., Aittomäki, K., Arola, J., Kruhoffer, M., Orntoft, T. F., Tomlinson, I. P., Kiuru, M., Arango, D. & Aaltonen, L. A. 2006. Distinct expression profile in fumarate-hydratase-deficient uterine fibroids. *Human Molecular Genetics* **15:** 97-103.

Venkatraman, E. S. & Olshen, A. B. 2007. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23:** 657-663.

Vissers, L. E., Veltman, J. A., van Kessel, A. G. & Brunner, H. G. 2005. Identification of disease genes by whole genome CGH arrays. *Human Molecular Genetics* **14 Spec No. 2:** R215-23.

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A.,Jr & Kinzler, K. W. 2013. Cancer genome landscapes. *Science* **339:** 1546-1558.

Watson, J. D. & Crick, F. H. 1953. The structure of DNA. *Cold Spring Harbor Symposia on Quantitative Biology* **18:** 123-131.

Wu, C., Orozco, C., Boyer, J., Leglise, M., Goodale, J., Batalov, S., Hodge, C. L., Haase, J., Janes, J., Huss, J. W.,3rd & Su, A. I. 2009. BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biology* **10:** R130-2009-10-11-r130. Epub 2009 Nov 17.

Wu, C., Wyatt, A. W., McPherson, A., Lin, D., McConeghy, B. J., Mo, F., Shukin, R., Lapuk, A. V., Jones, S. J., Zhao, Y., Marra, M. A., Gleave, M. E., Volik, S. V., Wang, Y., Sahinalp, S. C. & Collins, C. C. 2012. Poly-gene fusion transcripts and chromothripsis in prostate cancer. *Genes, Chromosomes & Cancer* **51:** 1144-1153.

Xu, Z. P., Wawrousek, E. F. & Piatigorsky, J. 2002. Transketolase haploinsufficiency reduces adipose tissue and female fertility in mice. *Molecular and Cellular Biology* **22:** 6142-6147.

Yan, Y. L., Tan, K. B. & Yeo, G. S. 2008. Right atrial isomerism: preponderance in Asian fetuses. Using the stomach-distance ratio as a possible diagnostic tool for prediction of right atrial isomerism. *Annals of the Academy of Medicine, Singapore* **37:** 906-912.

Zhang, F., Carvalho, C. M. & Lupski, J. R. 2009. Complex human chromosomal and genomic rearrangements. *Trends in Genetics* **25:** 298-307.